

# UTILITY PATENT APPLICATION TRANSMITTAL (Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No.  
EN999121

Total Pages in this Submission  
3

## TO THE ASSISTANT COMMISSIONER FOR PATENTS

Box Patent Application  
Washington, D.C. 20231

Transmitted herewith for filing under 35 U.S.C. 111(a) and 37 C.F.R. 1.53(b) is a new utility patent application for invention entitled:

**METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING DATA  
MOVEMENT WITHIN A COMPUTING ENVIRONMENT**

and invented by:

**Scott Thomas Marcotte**

If a CONTINUATION APPLICATION, check appropriate box and supply the requisite information:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No.: \_\_\_\_\_

Which is a:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No.: \_\_\_\_\_

Which is a:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No.: \_\_\_\_\_

Enclosed are:

### Application Elements

1. ☒ Filing fee as calculated and transmitted as described below
2. ☒ Specification having 57 pages and including the following:
  - a. ☒ Descriptive Title of the Invention
  - b. ☒ Cross References to Related Applications (if applicable)
  - c. ☐ Statement Regarding Federally-sponsored Research/Development (if applicable)
  - d. ☐ Reference to Microfiche Appendix (if applicable)
  - e. ☒ Background of the Invention
  - f. ☒ Brief Summary of the Invention
  - g. ☒ Brief Description of the Drawings (if drawings filed)
  - h. ☒ Detailed Description
  - i. ☒ Claim(s) as Classified Below
  - j. ☒ Abstract of the Disclosure

# UTILITY PATENT APPLICATION TRANSMITTAL (Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No.  
EN999121

Total Pages in this Submission  
3

## Application Elements (Continued)

3. ☒ Drawing(s) (when necessary as prescribed by 35 USC 113)

- a. ☒ Formal                      Number of Sheets                      17
- b. ☐ Informal                      Number of Sheets                      \_\_\_\_\_

4. ☒ Oath or Declaration

- a. ☒ Newly executed (original or copy)                      ☐ Unexecuted
- b. ☐ Copy from a prior application (37 CFR 1.63(d)) (for continuation/divisional application only)
- c. ☒ With Power of Attorney                      ☐ Without Power of Attorney
- d. ☐ DELETION OF INVENTOR(S)

Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. 1.63(d)(2) and 1.33(b).

5. ☐ Incorporation By Reference (usable if Box 4b is checked)

The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.

6. ☐ Computer Program in Microfiche (Appendix)

7. ☐ Nucleotide and/or Amino Acid Sequence Submission (if applicable, all must be included)

- a. ☐ Paper Copy
- b. ☐ Computer Readable Copy (identical to computer copy)
- c. ☐ Statement Verifying Identical Paper and Computer Readable Copy

## Accompanying Application Parts

8. ☒ Assignment Papers (cover sheet & document(s))

9. ☐ 37 CFR 3.73(B) Statement (when there is an assignee)

10. ☐ English Translation Document (if applicable)

11. ☒ Information Disclosure Statement/PTO-1449                      ☒ Copies of IDS Citations

12. ☐ Preliminary Amendment

13. ☒ Acknowledgment postcard

14. ☒ Certificate of Mailing

☐ First Class                      ☒ Express Mail (Specify Label No.):                      EL172581532US

# UTILITY PATENT APPLICATION TRANSMITTAL (Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No.  
EN999121

Total Pages in this Submission  
3

## Accompanying Application Parts (Continued)

15. ☐ Certified Copy of Priority Document(s) (if foreign priority is claimed)

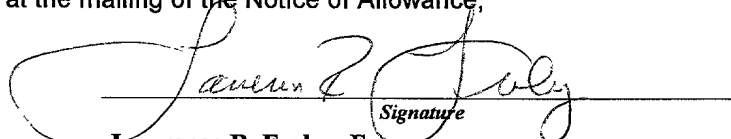
16. ☐ Additional Enclosures (please identify below):

## Fee Calculation and Transmittal

### CLAIMS AS FILED

For	#Filed	#Allowed	#Extra	Rate	Fee
Total Claims	45	- 20 =	25	x \$18.00	\$450.00
Indep. Claims	9	- 3 =	6	x \$78.00	\$468.00
Multiple Dependent Claims (check if applicable) <input type="checkbox"/>					\$0.00
BASIC FEE					\$760.00
OTHER FEE (specify purpose)					\$0.00
TOTAL FILING FEE					\$1,678.00

- ☐ A check in the amount of \_\_\_\_\_ to cover the filing fee is enclosed.
- ☒ The Commissioner is hereby authorized to charge and credit Deposit Account No. 09-0457(IBM) as described below. A duplicate copy of this sheet is enclosed.
- ☒ Charge the amount of \$1,678.00 as filing fee.
  - ☒ Credit any overpayment.
  - ☒ Charge any additional filing fees required under 37 C.F.R. 1.16 and 1.17.
  - ☐ Charge the issue fee set in 37 C.F.R. 1.18 at the mailing of the Notice of Allowance, pursuant to 37 C.F.R. 1.311(b).

  
Signature

Lawrence R. Fraley, Esq.  
Reg. No. 26,885  
IBM Corporation  
Intellectual Property Law N50/040-4  
1701 North Street  
Endicott, NY 13760-5553  
Telephone (607) 755-3207  
Facsimile (607) 755-3250

Dated: Nov 18, 1999

cc:



## CERTIFICATE OF MAILING BY "EXPRESS MAIL"

In Re Application of: Scott Thomas Marcotte

Title: METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING DATA  
MOVEMENT WITHIN A COMPUTING ENVIRONMENT

Attorney Docket No.: EN999121

"EXPRESS MAIL" MAILING LABEL NO. EL172581532US

Date of Deposit 11/18/99

I hereby certify that this paper is being deposited with the U.S. Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and addressed to Assistant Commissioner for Patents, Box PATENT APPLICATION, Washington, D.C. 20231.

Enclosed: New Utility Patent Application Transmittal  
Letter (Large Entity) (3 pp.) (in duplicate)  
U.S. Patent Application -  
Specification (40 pp.); Claims (16 pp.);  
Abstract (1 p.)  
Formal Drawings (17 sheets)  
Declaration and Power of Attorney (3 pp.)  
(     unsigned) (X signed)  
Assignment w/Recordation Cover Sheet (2 pp.)  
Information Disclosure Statement (2 pp.)  
Information Disclosure Citation w/references  
(2 pp.) (9 cited)  
2 Postcards

Denise M. Jurik

(Typed or printed name of person mailing paper or fee)

*Denise M. Jurik*  
(Signature of person mailing paper or fee)

**APPLICATION  
FOR  
UNITED STATES LETTERS PATENT**

**APPLICANT(S) NAME: Scott T. Marcotte**

**TITLE: METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING  
DATA MOVEMENT WITHIN A COMPUTING ENVIRONMENT**

**DOCKET NO. EN999121**

**INTERNATIONAL BUSINESS MACHINES CORPORATION**

**Certificate of Mailing Under 37 CFR 1.10**

I hereby certify that, on the date shown below, this correspondence is being deposited with the United States Postal Service in an envelope addressed to the Assistant Commissioner for Patents, Washington, D.C., 20231 as "Express Mail Post Office to Addressee".

"Express Mail" Label Number EL172581532US

On 11/18/99

Denise M. Jurik

*Typed or Printed Name of Person Mailing Correspondence*

*Denise M. Jurik*  
Signature of Person Mailing Correspondence

11/18/99  
Date

**METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING  
DATA MOVEMENT WITHIN A COMPUTING ENVIRONMENT**

**Cross-Reference to Related Applications**

5 This application contains subject matter which is  
related to the subject matter of the following application,  
which is assigned to the same assignee as this application  
and filed on the same day as this application. The below  
listed application is hereby incorporated herein by  
reference in its entirety:

10 "Managing The Flow Of Information Between Senders And  
Receivers Of A Computing Environment To Enhance Performance  
Of The Environment", Scott T. Marcotte, (Docket No.  
EN999066), Serial No. \_\_\_\_\_, filed herewith.

**Technical Field**

15 This invention relates, in general, to processing read  
and write requests within a computing environment, and in  
particular, to enhancing the performance of those requests.

**Background Art**

20 Enhancing the performance of read and write requests is  
an important goal of many computing systems, including  
distributed file systems, such as Distributed File Services

(DFS) systems. When writing or reading files, especially large files, the time it takes for the request and data to be transmitted across the network often adversely dominates the response time of the request. Thus, efforts have been made to reduce the network time and hence, the response time.

These efforts include locally caching data to avoid communications across the network and processing requests in parallel to improve response time. However, even with these efforts, systems still encounter considerable delays in data transmission, which negatively affects response time.

In addition to the above, response time is also negatively impacted by extra data movements within a computing system. For example, in systems, such as Distributed File Services (DFS) systems, data is moved from one set of buffers within the server to another set of buffers within the server. This data movement results in extra processing time at the server, which negatively impacts response time.

Based on the foregoing, a need still exists for a capability that enhances the performance of read and write requests. A further need exists for a capability that improves data transmission. A yet further need exists for a capability that eliminates extra data movements in an effort to enhance response time.

### Summary of the Invention

The shortcomings of the prior art are overcome and additional advantages are provided through the provision of a method of reducing data movement within a computing environment. In one embodiment, the method includes transmitting data between a file system of a computing unit of the computing environment and a transmission medium of the computing environment, wherein the transmitting includes bypassing non-file system buffers of the computing unit in performing the transmission.

In one example, the transmitting includes sending data from a sender of the computing environment over the transmission medium to the file system to be written to one or more storage media coupled to the file system. As a further example, the sending includes sending the data over the transmission medium using one or more buffers associated with the sender. The file system receives the data, wherein the receiving includes swapping one or more buffers associated with the file system with the one or more buffers associated with the sender.

In another example, the transmitting includes sending data from the file system over the transmission medium to a receiver of the data. In a further example, the sending includes using a routine identified by the receiver to send the data, wherein the routine is provided one or more pointers to the data to be sent to the receiver.



In a further aspect of the present invention, a method of reducing data movement within a computing environment is provided. The method includes, for instance, sending data from a sender of the computing environment to a file system of the computing environment, wherein the sending includes using one or more buffers associated with the sender; and receiving by the file system the data, wherein the receiving includes swapping one or more buffers associated with the file system with the one or more buffers associated with the sender.

In a further aspect of the present invention, a method of translating data from one format to another format is provided. The method includes, for instance, determining that data located in at least one buffer associated with a file system usable in writing data to one or more storage media coupled to the file system is to be translated from one format to another format; and translating at least a portion of the data in the at least one buffer. The translating being performed within the at least one buffer associated with the file system without requiring copying of the at least portion of the data to one or more other buffers.

Systems and computer program products corresponding to the above-summarized methods are also described and claimed herein.

Additional features and advantages are realized through the techniques of the present invention. Other embodiments

and aspects of the invention are described in detail herein and are considered a part of the claimed invention.

### **Brief Description of the Drawings**

5 The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the invention are apparent from the following detailed description taken in conjunction with the accompanying  
10 drawings in which:

FIG. 1 depicts one example of a computing environment incorporating and using the various aspects of the present invention;

15 FIG. 2 depicts further details of a computing unit of FIG. 1, in accordance with the principles of the present invention;

FIGs. 3a-3d depict one embodiment of the logic associated with processing requests, in accordance with the principles of the present invention;

20 FIG. 4 depicts one example of a data structure used in accordance with the principles of the present invention;

FIG. 5 depicts one embodiment of the logic associated with adding an operation to the data structure of FIG. 4, in accordance with the principles of the present invention;

5           FIG. 6 depicts one embodiment of the logic associated with allocating a buffer, in accordance with the principles of the present invention;

10           FIG. 7a depicts one example of an I/O vector used in accordance with the principles of the present invention;

            FIG. 7b depicts one example of a segment referencing a page, in accordance with the principles of the present invention;

15           FIG. 7c depicts one example of a segment table used in accordance with the principles of the present invention;

20           FIG. 8 depicts one embodiment of the logic associated with determining if a particular file operation can be performed, in accordance with the principles of the present invention;

            FIGs. 9a-9b depict one embodiment of the logic associated with a read operation, in accordance with the principles of the present invention;

FIGs. 10a-10b depict one embodiment of the logic associated with a write operation, in accordance with the principles of the present invention;

5 FIG. 11 depicts one embodiment of the logic associated with copying data, in accordance with the principles of the present invention;

FIG. 12 depicts one embodiment of the logic associated with ending a file operation, in accordance with the principles of the present invention;

10 FIG. 13 depicts one embodiment of the logic associated with removing a file operation from the data structure of FIG. 4, in accordance with the principles of the present invention; and

15 FIG. 14 depicts one embodiment of the logic associated with freeing a buffer, in accordance with the principles of the present invention.

### Best Mode for Carrying out the Invention

In accordance with one aspect of the present invention, a quick reply is sent from a receiver of a packet to a sender of the packet indicating receipt of the packet, prior to receiving the entire packet. This allows data to be sent to the receiver as fast as possible by, for example, keeping the data streaming over the communications or transmission medium between the sender and the receiver. In a further aspect of the present invention, a reply to a request is sent from a receiver of the request to the sender of the request, prior to providing the request to a file system coupled to the receiver.

In accordance with another aspect of the present invention, data movement within a computing environment is reduced, at the very least. In particular, data is transmitted between a file system of a computing unit of the computing environment and a transmission medium of the computing environment, such that non-file system buffers of the computing unit are bypassed in performing the transmission. As one example, during a file read operation, data is transmitted directly from the file system cache buffers to the requester of that data. As a further example, for a file write operation, buffer swapping in the file server is used to avoid data movement.

One embodiment of a computing environment incorporating and using the capabilities of the present invention is described with reference to FIG. 1. A computing environment

100 includes, for instance, at least one computing unit 102  
coupled to one or more other computing units 104. In one  
example, computing unit 102 is a server, while computing  
units 104 are clients. Each unit includes, for example, one  
5 or more central processing units, memory and one or more  
input/output devices, as is well known in the art.

Computing unit 102 is based, for instance, on the  
Enterprise Systems Architecture (ESA)/390 offered by  
International Business Machines Corporation, Armonk, New  
10 York. ESA/390 is described in an IBM publication entitled  
"Enterprise Systems Architecture/390 Principles of  
Operation," IBM Publication No. SA22-7201-04, June 1997,  
which is hereby incorporated herein by reference in its  
entirety. One example of a computing unit based on ESA/390  
15 is the 9672 Parallel Enterprise Server offered by  
International Business Machines Corporation.

One or more of computing units 104 are personal  
computers. As one example, a computing unit 104 is a  
personal computer executing Microsoft Windows, which runs on  
20 the Intel PC architecture. In one instance, one or more of  
computing units 104 include Server Message Block  
(SMB)/Common Internet File System (CIFS) clients.

Computing unit 102 is coupled to one or more of  
computing units 104 via a standard connection, such as any  
25 type of wire connection, token ring or network connection,  
to name just a few examples. One communications protocol  
used by one or more of these connections is TCP/IP.

The above-described computing environment and/or computing units are only offered as examples. The present invention can be incorporated and used with many types of computing units, computers, processors, nodes, systems, workstations and/or environments without departing from the spirit of the present invention. For example, one or more of the units may be based on the UNIX architecture or may include the Intel PC architecture. Additionally, while some of the embodiments described herein are discussed in relation to servers and clients, and in particular, to a file server and clients, such embodiments are only examples. Other types of receivers and senders of information, other types of servers and/or other types of computing environments can benefit from the present invention and are thus, considered a part of the present invention.

Additionally, in various aspects of the present invention, the clients need not be remote from the server. Various aspects of the invention are equally applicable to clients and servers running on the same physical machine, different physical machines or any combination thereof.

Further details of computing unit 102 are described with reference to FIG. 2. In one example, computing unit 102 includes an operating system 202, such as the OS/390 or MVS operating system offered by International Business Machines Corporation. Running on the operating system is, for instance, a file server 204. In one example, file server 204 is an OS/390 Unix Distributed File Services (DFS) server. File server 204 includes a plurality of layers,

such as, for instance, a physical file system (PFS) layer 206, a cache 208, a file specific server (filess) layer 210, an SMBparser layer 212, a Netbios layer 214 and an Asynchronous Socket I/O layer 216. Although each of the  
5 layers is shown as a part of the file server, one or more of the layers can be coupled to the file server. Thus, the layers are considered associated with the file server, as well as with the computing unit.

Physical file system 206 is the layer of the file  
10 server used to write information to and/or read information from the storage media (such as, for instance, disks) coupled to the physical file system. In one example, the physical file system retrieves information from its cache 208 and writes that information to the storage media. As a  
15 further example, the physical file system retrieves information from the storage media, places it in its cache, and then forwards it to the requesters of that information.

File specific server layer 210 is the layer that transforms SMBparser requests into cache and physical file  
20 system requests, and also interfaces with a DFS token management service to manage the sharing of files between clients.

SMBparser layer 212 is the main SMB processing layer  
25 that knows the state of the environment relative to the client (e.g., what files are open, etc.). When this layer is called, the client session is in a stopped state. Thus, no more requests are received from the client, while it is



in this state. The lowest SMB layers call Netbios to re-enable the session after performing some preliminary processing.

Netbios layer 214 is responsible for maintaining  
5 communications between the server and the clients. It is  
conduit between the SMBparser layer and the Asynchronous  
sockets layer. Netbios schedules the asynchronous receive  
requests, described below, on behalf of SMBparser. For  
example, SMBs are packaged in Netbios packets and Netbios  
10 makes the asynchronous socket calls on behalf of the  
SMBparser layer, which performs the work.

Async sockets I/O layer 216 provides the low level  
socket communications that maintain the thread pools used in  
processing client requests. In one example, the  
15 Asynchronous sockets layer uses the POSIX (Portable  
Operating System Interface for Computer Environments)  
Asynchronous I/O interface to handle communications.

In one aspect of the present invention, quick replies  
are used to enable a reply to a request be sent, prior to  
20 the receiver (e.g., a server) receiving the entire request  
from a sender (e.g., a client). Further, in another  
example, quick replies are used to send a reply to a request  
(either a partially or fully received request), prior to  
providing the request to a file system that is to process  
25 the request (e.g., write the data of the request to disk).  
In another aspect of the present invention, data movement  
within the server is eliminated or at the very least

reduced, when transmitting data between the file system associated with the server and a transmission medium (e.g., a wire) between the server and the client, by bypassing non-file system buffers during the transmission. These aspects of the present invention, as well as others, are described in detail with reference to the overview logic depicted in FIGs. 3a-3d. (These figures depict only one example. Various modifications, additions and deletions can be made without departing from the spirit of the present invention. Additionally, although this example is described with reference to a server and a client, the invention is not limited to the same.)

Referring to FIG. 3a, at some point in time, the server is notified by POSIX Asynchronous I/O Services that information (e.g., a new request) has arrived from a client, STEP 300. This information is transmitted, for instance, in a packet. However, in accordance with one aspect of the present invention, only a portion of the packet is received by the lowest communications layers of the server (e.g., Async Sockets I/O and Netbios). For example, Netbios receives 12 bytes of information: 4 bytes are the Netbios header and 8 bytes are the beginning of a Server Message Block (SMB) packet (i.e., an SMB header). Netbios strips off the Netbios header and passes the request onto a processing thread to handle the SMB packet. In particular, Netbios notifies SMBparser that a new request has arrived. This allows SMBparser to control when the packet is fully consumed from the transmission medium, when a reply is sent

for the SMB, and when the client session should be enabled to receive the next packet.

Since SMBparser is notified when at least one byte of data has been received, the server initially ensures that at least the Netbios header has been received before continuing processing. Thus, a determination is made as to whether a predefined number of bytes (X) has been received, INQUIRY 302. In this example, a determination is made as to whether 4 bytes of data, representing the Netbios header, have been received. If it is determined that the Netbios header has not been received, then the server waits until the predetermined number of bytes have been received via a synchronous read call, STEP 304, and processing continues with INQUIRY 302.

On the other hand, when the predefined number of bytes have been received, then a further determination is made as to whether the request represents a data packet (e.g., an SMB packet), INQUIRY 306. This is determined by examining the Netbios header. If the request is not a data packet and is instead, for example, a Netbios packet, then another determination is made as to whether the request is less than N bytes, where N is equal to 12, in this example, INQUIRY 308.

If the request is less than N bytes, then the request (which is equal to 4 bytes in this example) is processed, STEP 312, and flow returns to INQUIRY 302 with the remaining bytes (e.g., 8 bytes) to be processed. However, if the

request is greater than N bytes, then the rest of the packet is received and processed, STEP 314. Thereafter, the socket is enabled to receive the next request asynchronously, STEP 316. In one example, the socket is enabled by issuing an asynchronous receive function that allows notification of when a client request arrives at the server. Processing is then complete for this request and thus, this thread becomes available for another request.

Returning to INQUIRY 306, if the received packet is a data request, then a determination is made as to whether a predetermined number of bytes (e.g., Y=8) of the SMB has been received, INQUIRY 320. If the first 8 bytes of the SMB header, in this example, have not been received, then the server waits for those bytes, STEP 322, and processing continues with STEP 320. On the other hand, if the predetermined number of bytes has been received, then the SMBparser layer is called to process the SMB, STEP 324. For example, the address of the predetermined number of bytes is passed to SMBparser with the socket in a stopped state. Since the socket is in the stopped state, no more information will be examined on the socket until specifically requested by the SMBparser layer.

One embodiment of the logic associated with processing the SMB is described with reference to FIG. 3b. Initially, the SMBparser layer examines the type of packet to determine whether the packet represents a write request or some other type of request, STEP 330 (FIG. 3b). In one example, this determination is made by examining the SMB header of the

request. If the request is not a write request, then the rest of the SMB is received from the client over the transmission medium via a synchronous TCP/IP receive function, STEP 332.

5           Subsequently, a determination is made as to whether the request is an open file based operation, INQUIRY 334. If it is an open file based operation, then the request is added to a data structure to ensure the request is processed in proper order, as explained below, STEP 336. One example of  
10   such a data structure is described with reference to FIG. 4.

          In one embodiment, a data structure 400 is a linked list (e.g., singly or doubly linked) representation of a queue having one or more op\_queue structures 402 and a file handle 404. Each op\_queue structure represents a pending  
15   operation, and includes a plurality of fields, such as the following: a next field indicating the next request on the queue; a type field indicating whether the operation is a read operation, a write operation, or another type of operation; a waiters field indicating if threads are waiting  
20   for the request to finish; and a done field which is a flag that indicates if the operation has completed its file system calls.

          File handle 404 represents a file that is opened for a particular client. The file handle includes various fields,  
25   including, for instance, the following: a lock field (file lock) which is used for serialization; a queueH field that points to the head of data structure 400 (e.g., the queue);

a queueT field that indicates the tail of the data structure; and an error field (fi\_error) that indicates the most recent error found when processing the file.

One embodiment of the logic associated with adding a file operation to the queue is described with reference to FIG. 5. In one example, an add routine (e.g., addNewFileOp(int type, int \*preverror, op\_queue \*cookie)) is used to add a file operation to the queue, return any prior error (preverror) from any previous operation, and return a cookie which can be used for future calls. In one example, the cookie is the address of the associated op\_queue structure. The routine is called when a new operation is received by SMBparser, but before a reply is sent to the client. It establishes the position of the file operation in the queue, but it does not block the service thread of the request.

The add routine receives as input the type of the request and proceeds to allocate and initialize an op\_queue structure, STEP 500. As part of the initialization, the type of the request is saved in the op\_queue, the rest of fields are cleared, and the \*cookie is set equal to the address of the op\_queue.

Thereafter, the file lock of the file handle is obtained exclusively, so that the data structure is not updated by another function during the add operation, STEP 502. Subsequently, the op\_queue structure is added to the end of the data structure queue and QueueT of the file

handle is updated, STEP 504. The exclusive lock is then released, STEP 506, and any prior error and the cookie are returned to SMBparser, STEP 508. This concludes the adding of an operation to the queue.

5           Returning to FIG. 3b, subsequent to adding the file operation to the file operation queue, STEP 336, socket communications is enabled for the next SMB, STEP 338. In one example, a POSIX asynchronous I/O receive function is used to enable the communications. Additionally, the file  
10 specific server is called to perform any additional processing and to make appropriate calls to the physical file system, as described below, STEP 340.

          Returning to INQUIRY 330, if the request is a write request (e.g., a write or writeraw request), then the rest  
15 of the SMB header is received via the synchronous TCP/IP receive function (size of header is known), STEP 342. At this point, no user data has been received for the write operation.

          Thereafter, the write file operation is added to the  
20 file operation queue in the manner described above, STEP 344, and a reply is sent back to the client via a synchronous TCP/IP writev function, STEP 346. In one example, this reply is a quick reply sent prior to receiving the entire packet.

25           Additionally, the server (e.g., SMBparser) gets ready to receive the data in the physical file system. In one

example, this includes having SMBparser obtain one or more physical file system buffers, based on the offset/length information in the SMB, STEP 348. In one instance, the buffers are correlated to pages (each page includes, for example, 4096 bytes) and are thus, obtained using a pfs\_allocate page routine. In other embodiments, the buffers do not correlate to pages, but instead, represent some other amount of memory.

One embodiment of the logic associated with an allocate page routine is described with reference to FIG. 6. Initially, a determination is made as to whether a free list indicating a list of free pages is empty, INQUIRY 600. If the free list is not empty, then a page is obtained from the list, STEP 602. Otherwise, a page is obtained from the program heap, STEP 604. Subsequent to obtaining a page, the page address is returned to the caller, STEP 606. This page address represents the address of the buffer to hold the data.

Returning to FIG. 3b, in addition to obtaining the buffers for the data, an I/O vector is created for page aligned data, STEP 349. One embodiment of such an I/O vector is described with reference to FIG. 7a.

An I/O vector 700 is, in one instance, an array having one or more elements 702. Each element includes, for example, a buffer pointer indicating the start of a buffer (e.g., an address into or the start of a page of memory), and a size reflecting the size of the buffer. The I/O



vector is created and initialized based on the buffers  
obtained in STEP 348.

The buffers are in the physical file system cache and,  
in this instance, are made up of one or more pages of  
5 memory. A page 710 (FIG. 7b) in the physical file system  
cache is pointed to by a data structure, referred to as a  
segment 712.

Each segment has an array of page pointers (e.g., 16  
pointers) used to locate one or more pages. The segment  
10 also has a status array used to indicate the status of a  
corresponding page, and a lock for serialization purposes.

Segments are located using segment tables 722 (FIG.  
7c). Each segment table has a number of fields including,  
for instance, a next field indicating the next segment  
15 table; a segfirst field indicating the segment number of the  
first segment of that segment table; a count field  
indicating the number of segments of the segment table; and  
a segment array including pointers to the one or more  
segments of the segment table.

20 The list of one or more segment tables is referenced by  
a segtable field of a physical file system file handle 724.  
The file handle also includes a lock, which is used for  
serialization.

Returning to FIG. 3b, after creating the I/O vector,  
25 STEP 349, the data is received via, for instance, the SMB

packet, STEP 350. For writeraw operations, more user data will flow from the client and that data is also received at the server at the same time. In accordance with one aspect of the present invention, the data is read directly into the PFS cache compatible buffers at page aligned offsets. That is, other buffers within the server are bypassed when receiving the data over the transmission medium, as described in more detail below.

In addition to receiving the data, the socket communications for the next SMB is enabled via an asynchronous receive call, STEP 338. This starts the socket session with the client which means the server will be notified when the next SMB request from the client is received (which could be dispatched and processed in parallel with the thread that is processing the SMB\_write or SMB\_writeraw request).

Sending the reply before receiving the rest of the data can greatly reduce the time before the next client request is received by the server because the client reply is on its way before the original packet is received, and for file writes that packet can be large (e.g., 64k for SMB/CIFS protocol) and take a significant amount of time. Thus, replies are seen sooner by the client, which can then send the next data sooner and hence, data is sent to the server faster. Therefore, the communications time (the main bottleneck) for file writes is reduced. Additionally, the queueing is performed on a file handle for a specific client, no other client requests are blocked to the file and

other requests from the same client that affect other files or directories are not blocked by the file queueing capabilities described herein allowing for maximum parallelism. The file handle is created when the file is  
5 opened (e.g., via an SMB\_open operation) and destroyed when the file is closed (e.g., via an SMB\_close operation).

Subsequent to enabling the socket communications, the file specific server is called to perform any additional processing and to make appropriate calls to the physical  
10 file system, STEP 340.

One embodiment of the processing associated with the file specific server and the physical file system is described with reference to FIG. 3c. As an example, when the file specific server is called, the file specific server  
15 performs any additional processing that needs or is desired to be performed, STEP 352. In one example, for OS/390 DFS, this includes token management and error checking. The quick reply aspect of the present invention allows this processing to be performed in parallel, if two or more  
20 client file operations for the same file have arrived at the server.

Thereafter, a determination is made as to whether this is an open file based operation, INQUIRY 354. If this is an open file based operation, a further inquiry is made as to  
25 whether the operation can continue, INQUIRY 356. As one example, the operation can continue if there are no operations in progress before it that block the operation.

This determination is made, for instance, by a start file operation. The start file operation is called before the physical file system is called for a particular operation. It is the routine that blocks the thread, if it is not next  
5 in the list of requests allowed to be processed by the physical file system.

In particular, the start file operation examines the file handle to see if the request is allowed to progress. For a write operation, it will progress if it is the next  
10 unprocessed request in the queue. For a read operation, it will progress if it is before a write operation. Read type request are allowed in parallel, so many read requests are allowed in parallel, if there are no unfinished write requests ahead of them in the queue. Thus, before the  
15 physical file system is called, a permission to call PFS function is called which checks to see if the thread is processing a request which is next in-line or one that is to wait its turn (in proper client receipt order). This allows two or more SMB requests that pertain to the same file to be  
20 processed in parallel, except for the actual physical file system call. Thus, parallel processing is not reduced.

One embodiment of a start file operation is described with reference to FIG. 8. The start file operation receives as input a cookie, which identifies the address of the  
25 op\_queue associated with this request. Initially, the lock of the file handle corresponding to the input op\_queue is obtained exclusively in order to prevent other updates to the queue during this processing, STEP 800. Thereafter, a

variable referred to as nextEntry is set equal to the first op\_queue in the queue, STEP 802.

5 A determination is then made as to whether nextEntry is equal to the input op\_queue, INQUIRY 804. If it is equal, then the lock is released, STEP 806, and the start operation is complete. This indicates that there are no operations in progress before this request that can block the operation.

10 On the other hand, if the nextEntry is not equal to the input op\_queue, INQUIRY 804, then a further determination is made as to whether the op\_queue (represented in nextEntry) is marked as complete, INQUIRY 808. If the op\_queue is not marked as complete, then yet another determination is made as to whether the incomplete op\_queue indicates the type is a write operation or whether the input op\_queue indicates a write operation, INQUIRY 810. If neither the incomplete op\_queue or the input op\_queue indicates a write, or if the op\_queue is marked as complete, then nextEntry is set equal to the next entry in the list, STEP 812. Processing then continues with INQUIRY 804.

20 However, if either the incomplete op\_queue or the input op\_queue indicates a write operation, then the waiters field in nextEntry (i.e., the prior blocking op\_queue, not the input op\_queue) is incremented by one, STEP 814. Thereafter, the lock is released, STEP 816, and the operation is put to sleep, waiting to be awoken, STEP 818.

Thus, the start operation makes the thread wait until there are no more operations in the queue ahead of it that are marked incomplete and not incompatible with the type (reads are compatible with reads, writes are compatible with nothing). When the routine returns, the thread is next in line to make the physical file system call. Thus, if there are blockers, the thread is put to sleep until blocking operations are complete. Then, processing continues with STEP 800.

Returning to FIG. 3c, and in particular, to INQUIRY 356, if the operation cannot continue, then the thread is put to sleep as indicated above, STEP 358, and sleeps until it is awoken. Additionally, a prior write error is returned, if any.

When the operation can continue, then a determination is made as to whether the file operation is a read operation, INQUIRY 360. This is determined by the type field in the op\_queue. If it is a read operation, then the physical file system is called to read the file data. In one example, a pfs\_read routine (file\_handle \*fh, int offset, int size, int \*resid, ...void (\*copyrtn)(), int cookie) is used to read the data. Input to the pfs\_read routine is a callback function address, a file offset, and an amount of data to be read. The callback function sends data directly over the transmission medium from the physical file system buffers to the requester of the data. Thus, additional buffers within the server are bypassed, when transmitting data from the physical file system to the

transmission medium. The physical file system calls the callback function with the addresses of the pfs cache buffers that contain the data, when data is in its buffers. One example of a physical file system read function is  
5 described with reference to FIGs. 9a-9b.

Referring to FIG. 9a, initially, an empty I/O vector is initialized, STEP 900. This vector is initialized to include pointers to any buffers used by the physical file system to hold the data that is to be transmitted over the  
10 transmission medium to the client issuing the request. Additionally, the lock in the physical file system file handle is locked in read mode in order to obtain information from the segment table, STEP 902.

Thereafter, the first segment and last segment numbers  
15 affected by the read operation are calculated, STEP 904, and a variable referred to as segment is set equal to the first segment, STEP 906.

Next, a determination is made as to whether the  
20 variable referred to as segment is greater than the last segment, INQUIRY 908. If segment is greater than the last segment, then no further segments need to be processed, and the lock in the pfs file handle is released, STEP 910.

However, if segment is not greater than the last  
25 segment, then the address of the segment is obtained based upon the segment number and segment table. Further, the

segment table and segment are obtained, if they do not exist in memory, STEP 912.

5       Thereafter, the lock field in the segment structure is set to read mode, STEP 914. After setting the lock in read mode, the first page and last page numbers affected by the read operation are calculated, STEP 916 (FIG. 9b). Additionally, a variable referred to as page is set equal to the first page, STEP 918.

10       Subsequently, a determination is made as to whether page is greater than last page, INQUIRY 920. Should page be greater than last page, then the callback function is called, STEP 922. In particular, the callback function is called with an I/O vector address and an indication if it will be the last time the function is to be called or an error occurs.

15

Subsequently, the segment lock is released, STEP 924, and the variable referred to as segment is incremented by one, STEP 926. Processing then continues with INQUIRY 908.

20       Returning to INQUIRY 920, if the variable referred to as page is not greater than the last page, then the address of the page is obtained based on the segment and page number, STEP 928. Additionally, the page or pages are read into storage, if they are not in memory. Further, there is a wait for I/O, if there is a pending read from disk or if

25       the page is pending write to disk and the data is marked as translated, STEP 928.



Subsequently, a determination is made as to whether the file system indicated data is translated, INQUIRY 930. In one example, this determination is made by checking a bit mask to see if the buffer is translated. If the data is indicated as translated, then the segment lock is released, STEP 932, and the segment lock is then obtained in write mode, STEP 934. Additionally, an untranslate function is called to untranslate the data, STEP 936. In accordance with one aspect of the present invention, the data translation (or untranslation) is handled internally in the physical file system in order to reduce data movement. In particular, the data is translated directly in the physical file system cache buffers. The data in the buffers are translated before calling the callback function.

After translating the data into the format of the client (e.g., into EBCDIC or ASCII), the segment lock is released, STEP 938, and the segment lock is then obtained in read mode, STEP 940.

Thereafter, or if the data is not translated, INQUIRY 930, then the address of the first byte in the page covered by the I/O operation is added to the I/O vector, along with the number of bytes covered by the I/O operation, STEP 942. Additionally, the variable referred to as page is incremented by one, STEP 944. Processing then continues with INQUIRY 920.

Returning to FIG. 3c, and in particular, to INQUIRY 360, if the request is not a read operation, then a further

determination is made as to whether the operation is a file write operation, INQUIRY 364. If it is a write operation, then a pfs\_write routine is called with the data, which is in the physical file system compatible buffers. The  
5 physical file system performs buffer swapping whenever possible, directly updating the caller's I/O vector with old buffers used for the file.

A pfs\_write routine (file\_handle \*fh, int offset, int size, struct iovec \*iov, int iovcnt, ..., int flag) receives  
10 as input an I/O vector of buffer addresses, a file handle, an amount of bytes of data to write, as well as a flag indicating if the incoming data is in physical file system compatible buffers. If the flag indicates compatible buffers, then the physical file system swaps its own  
15 internal buffers with the caller's buffers. The IOV field points to a vector of buffer addresses and lengths (the buffers need not be contiguous). An IOV count indicates how many buffers are provided. However, if a write is not on a buffer boundary (i.e., the offset is not evenly divisible by  
20 the buffer size, then some copying in the first and last buffer in the list may be necessary).

By swapping the buffers, data movement is eliminated or at least substantially reduced. For example, when the physical file system is called, it can perform buffer  
25 swapping, which usually means less data movement when the file lock is held, which can also reduce lock contention. The physical file system updates the caller's buffer vector with the old pfs cache buffers that the pfs originally held

for the file. Then, when SMBparser completes the write request, it frees the buffers; although, the buffers may not be the same as the buffers requested. The cache buffers are written to disk when the physical file system determines it is time or by a later request (such as, for example, an fsync request, described below). By swapping the buffers, rather than reading the data directly from the transmission medium into the physical file systems buffers, an exclusive lock on the file or portion of the file being processed is not held across expensive receives from the transmission medium.

One embodiment of a write routine is described with reference to FIGs. 10a-10b. Initially, the physical file system file handle lock is locked in write mode, STEP 1000. Thereafter, the first segment and last segment numbers affected by the write operation are calculated, STEP 1002, and a variable referred to as segment is set equal to the first segment, STEP 1004.

Next, a determination is made as to whether the variable referred to as segment is greater than the last segment, INQUIRY 1005. If segment is greater than the last segment, then no further segments need to be processed, and the lock in the pfs file handle is released, STEP 1006.

However, if segment is not greater than the last segment, then the address of the segment is obtained based upon the segment number and segment table. Further, the

segment table and segment are obtained, if they do not exist in memory, STEP 1008.

Thereafter, the lock field in the segment structure is set to write mode, STEP 1010. After setting the lock in write mode, the first page and last page numbers affected by the write operation are calculated, STEP 1012. Additionally, a variable referred to as page is set equal to the first page, STEP 1013.

Subsequently, a determination is made as to whether page is greater than last page, INQUIRY 1014. Should the page be greater than the last page, then a determination is made as to whether all the pages in the segment are dirty, INQUIRY 1016. If so, then an I/O operation is scheduled to write the data to disk, STEP 1018. Thereafter, or if the data is not dirty, then an indication is made by the file system as to whether the data is translated, STEP 1020. This indication is saved in the status array for the page.

Subsequently, the segment lock is released, STEP 1022, and the variable referred to as segment is incremented by one, STEP 1024. Processing then continues with INQUIRY 1005.

Returning to INQUIRY 1014, if the variable referred to as page is not greater than the last page, then the address of the page is obtained based on the segment and page number, STEP 1026. Additionally, the page or pages are read into storage, if they are not in memory and the whole page

is not to be fully updated. Further, there is a wait for I/O, if there is a pending I/O operation, STEP 1026.

Subsequently, a determination is made as to whether the caller is updating the entire page with the request, INQUIRY 1028. If so, then the buffer address in the input I/O vector is switched with the page address, STEP 1030, and the status of the page is marked as untranslated, STEP 1032. Additionally, the status of the page is marked as dirty, STEP 1036, and the variable referred to as page is incremented by one, STEP 1038. Processing then continues with INQUIRY 1014.

Returning to INQUIRY 1028 (FIG. 10b), if the caller is not filling an entire buffer, then data is copied, STEP 1034. In one example, the data is copied using a copy routine, one embodiment of which is described with reference to FIG. 11.

Initially, a determination is made as to whether the page is translated, INQUIRY 1100. This determination is made by checking the status field. If it is translated, then the parts not updated by this write operation are untranslated, STEP 1102.

After translating or if translation is not necessary, then the data is copied to the page, STEP 1104. This completes the copy routine, and processing then continues with STEP 1036 of FIG. 10b.

Returning to FIG. 3c, and in particular, INQUIRY 364, if the request is not a write file operation, then a further determination is made as to whether it is a close operation, INQUIRY 368. If it is a close operation, then it is the  
5 last operation to be performed for the file. Thus, all data is written to the pfs by close time and the file close sees any errors that happened after any quick reply of a write request. Thus, an I/O operation is scheduled to write the data to disk, STEP 370. In one example, a pfs\_schedule  
10 function is called to perform the scheduling. As one example, the scheduling logic loops through all the segments for the file that have dirty pages and schedules an I/O operation for each dirty segment.

Next, a determination is made as to whether write-through is requested, INQUIRY 372. If write-through is  
15 requested, then I/O operations are scheduled, as described above; however, processing waits until the I/O operations are complete, STEP 374. If this was not a write-through request, then waiting is not necessary.

Returning to INQUIRY 368, if this is not a close  
20 operation, then the physical file system is called to perform the specified operation, STEP 376.

Subsequent to performing the requested operation, STEPS 362, 366, 372, 374 or 376, processing continues with INQUIRY  
25 380 of FIG. 3d. For example, a determination is made as to whether this is an open file based operation, INQUIRY 380. If it is an open file based operation, then an indication is

made that the operation is done and any sleepers are awoken,  
STEP 382. In one example, a stop file routine is used in  
order to perform this processing. This routine is called  
after the physical file system is called. It marks the  
5 operation as complete, but does not remove it from the  
queue. It does wake up any waiting threads, so that they  
can proceed to calling the physical file system on behalf of  
their operation. It stores the error code (if any), if the  
physical file system request failed, so other requests can  
10 detect the error. This allows prior errors from previous  
writes to be reflected to later writes, so client  
applications see the errors, even if the error is for a  
prior write operation.

One embodiment of the logic associated with the stop  
15 file operation is described with reference to FIG. 12. As  
one example, input to the stop file routine is the cookie  
having the op\_queue address and the error indicator  
including the error from the physical file system, if any.

Referring to FIG. 12, initially, the file handle lock  
20 is obtained for serialization purposes, STEP 1200, and the  
Done flag in the input op\_queue is set to complete, STEP  
1202. Additionally, the fi\_error of the file handle is set  
equal to the input error (if any), STEP 1204.

Subsequently, a determination is made as to whether the  
25 waiters field in the input op\_queue is greater than zero,  
INQUIRY 1206. If it is greater than zero, then the waiters  
are awoken, STEP 1208, and the waiters count is set equal to

zero, STEP 1210. Thereafter, or if the waiters field is not greater than zero, then the lock is released, STEP 1212. This completes the stop operation used to indicate that the operation is complete and to wake up the waiters.

5           Returning to FIG. 3d, after indicating that the operation is done or if this is not an open file based operation, any additional processing is performed, if necessary or desired, STEP 384. This processing is allowed to be performed in parallel for multiple operations hitting  
10       the same file.

              Subsequently, a determination is made as to whether this is a file close operation, INQUIRY 385. If this is not a file close operation, then SMBparser sends a reply SMB to the client, if one has not already been sent, STEP 386. On  
15       the other hand, if this is a file close operation, then a determination is made as to whether there are any previous errors, STEP 387. In one example, this determination is made by a query routine that retrieves the latest error on the file, if needed or desired by SMBparser. In particular,  
20       the routine returns the value stored in the fi\_error field. Input to the routine is the op\_queue cookie, and the perror field.

              Subsequent to determining if there are any errors, the reply is sent, if it has not already been sent, STEP 386.  
25       This reply will include any error determined in the previous step.



In addition to the above, a remove operation is performed, STEP 388. In one example, this routine removes the file operation from the queue and records any additional error found after calling the physical file system.

5 Returned from this routine is any prior error found. When the SMB close operation calls the remove file operation, any previous physical file system call errors for the file will be included. Thus, the client is informed in its reply, if an error occurred.

10 One embodiment of a remove operation is described with reference to FIG. 13. As one example, input to the remove operation is a cookie having the address of the op\_queue; any post physical file system error (error); and an address to return any prior error (perror).

15 Referring to FIG. 13, initially, a file handle lock is obtained, STEP 1300, and the file error (fi-error) is set equal to the error, if any, STEP 1302. Additionally, the input op\_queue is removed from the queue data structure, STEP 1304.

20 Thereafter, a determination is made as to whether the waiters field in the input op\_queue is greater than zero, INQUIRY 1306. If waiters is greater than zero, then the waiters are awoken, STEP 1308. Additionally, the waiters count is set to zero, STEP 1310, and perror is set equal to  
25 the fi\_error, STEP 1312.

Subsequently, or if waiters is not greater than zero, then the file handle lock is released, STEP 1314, and the cookie is set equal to zero, STEP 1316. This concludes the remove operation.

5           Returning to FIG. 3d, in addition to performing the remove operation, the data buffers are returned to free storage, STEP 390. In one example, the data buffers are returned via a free page routine, one embodiment of which is described with reference to FIG. 14.

10           As one example, in order to return each data buffer, the page is added to the front of the free page list, STEP 1400. This is done for each data buffer to be returned. Often, the data buffers that are returned are not the original buffers. This completes the processing associated  
15 with various aspects of the present invention.

Described in detail above are capabilities that allow data to be transmitted directly from the physical file system cache buffers to the client for file reads; the ability to use buffer swapping for file writes to avoid data  
20 movement, especially for large file writes, in one instance; the ability to use hand-shaking and partial packet receive in the lower communications and SMBparsr layers to allow for quick replies for file writes, as one example, reducing delays in data transmission; and the ability to provide file  
25 operation queueing and latent error handling, which ensures the integrity of file writes given that quick replies are used.

One or more aspects of the present invention advantageously increase throughput as viewed by the end-user and also reduces server processor usage. Reducing the data movement in the server reduces server processor time per  
5 file read/write operations, and the quick replies ensure that data is sent to the server as fast as possible for file writes keeping the data streaming on the transmission medium.

10 With the various aspects of the present invention, no advance notice or prediction on the access pattern of data in the files is necessary. Additionally, one or more aspects of the present invention are independent of the communications adapter used.

15 The present invention is not limited to file serving or to SMB/CIFS filing serving. For example, the data movement reduction technique can be applied for any application, be it a file server or just an application running on the local machine that is calling the physical file system to reduce data movement.

20 As a further example, the partial packet receive processing and the quick reply techniques are applicable to various servers that processes requests from clients. For instance, they are applicable to a server that processes requests, where those requests affect some object the server  
25 manages, and the client requests that affect the object have a "start using object" and a "stop using object" request that bound the use of the object. An unlimited number of

requests could occur between the start/stop request. For the file server, the start operation is file open and the stop operation is a file close.

5 The present invention can be included in an article of manufacture (e.g., one or more computer program products) having, for instance, computer usable media. The media has embodied therein, for instance, computer readable program code means for providing and facilitating the capabilities of the present invention. The article of manufacture can be  
10 included as a part of a computer system or sold separately.

Additionally, at least one program storage device readable by a machine, tangibly embodying at least one program of instructions executable by the machine to perform the capabilities of the present invention can be provided.

15 The flow diagrams depicted herein are just exemplary. There may be many variations to these diagrams or the steps (or operations) described therein without departing from the spirit of the invention. For instance, the steps may be performed in a differing order, or steps may be added,  
20 deleted or modified. All of these variations are considered a part of the claimed invention.

Although preferred embodiments have been depicted and described in detail herein, it will be apparent to those skilled in the relevant art that various modifications,  
25 additions, substitutions and the like can be made without

departing from the spirit of the invention and these are therefore considered to be within the scope of the invention as defined in the following claims.

400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

## Claims

What is claimed is:

- 1           1.    A method of reducing data movement within a  
2    computing environment, said method comprising:  
  
3                transmitting data between a file system of a  
4                computing unit of said computing environment and a  
5                transmission medium of said computing environment; and  
  
6                wherein said transmitting comprises bypassing non-  
7                file system buffers of said computing unit in  
8                performing the transmission.
- 1           2.    The method of claim 1, wherein said transmitting  
2    comprises sending data from a sender of said computing  
3    environment over said transmission medium to said file  
4    system to be written to one or more storage media coupled to  
5    said file system.
- 1           3.    The method of claim 2, wherein said sending  
2    comprises sending said data over said transmission medium  
3    using one or more buffers associated with said sender, and  
4    wherein said method further comprises receiving said data by  
5    said file system, said receiving comprising swapping one or  
6    more buffers associated with said file system with said one  
7    or more buffers associated with said sender.

1           4. The method of claim 3, wherein said swapping avoids  
2 copying data from said one or more buffers associated with  
3 said sender to said one or more buffers associated with said  
4 file system.

1           5. The method of claim 3, further comprising obtaining  
2 said one or more buffers associated with said file system,  
3 prior to said receiving.

1           6. The method of claim 3, further comprising returning  
2 said one or more buffers associated with said file system to  
3 free storage, subsequent to said swapping.

1           7. The method of claim 3, further comprising  
2 translating data in said one or more buffers associated with  
3 said file system to a format compatible with said file  
4 system, said translating being performed within said one or  
5 more buffers associated with said file system.

1           8. The method of claim 7, further comprising  
2 determining that translating is to be performed, said  
3 determining comprising checking a translation indicator to  
4 determine if translation is to be performed.

1           9. The method of claim 1, wherein said transmitting  
2 comprises sending data from said file system over said  
3 transmission medium to a receiver of said data.

1           10. The method of claim 9, wherein said sending  
2 comprises using a routine identified by said receiver to  
3 send said data, wherein said routine is provided one or more  
4 pointers to said data to be sent to said receiver.

1           11. The method of claim 9, wherein said sending  
2 comprises providing to said receiver one or more pointers to  
3 said data.



1           12. A method of reducing data movement within a  
2 computing environment, said method comprising:

3                 sending data from a sender of said computing  
4 environment to a file system of said computing  
5 environment, said sending comprising using one or more  
6 buffers associated with said sender; and

7                 receiving by said file system said data, wherein  
8 said receiving comprises swapping one or more buffers  
9 associated with said file system with said one or more  
10 buffers associated with said sender.

1           13. The method of claim 12, wherein said swapping  
2 avoids copying data from said one or more buffers associated  
3 with said sender to said one or more buffers associated with  
4 said file system.

1           14. A method of translating data from one format to  
2 another format, said method comprising:

3                 determining that data located in at least one  
4 buffer associated with a file system usable in writing  
5 data to one or more storage media coupled to said file  
6 system is to be translated from one format to another  
7 format; and

8                 translating at least a portion of said data in  
9 said at least one buffer, said translating being  
10 performed within said at least one buffer associated  
11 with said file system without requiring copying of said  
12 at least portion of said data to one or more other  
13 buffers.

1           15. The method of claim 14, wherein said determining  
2 comprises checking a translation indicator to determine if  
3 translation is to be performed.

1           16. A system of reducing data movement within a  
2 computing environment, said system comprising:

3           means for transmitting data between a file system  
4 of a computing unit of said computing environment and a  
5 transmission medium of said computing environment; and

6           wherein said means for transmitting comprises  
7 means for bypassing non-file system buffers of said  
8 computing unit in performing the transmission.

1           17. The system of claim 16, wherein said means for  
2 transmitting comprises means for sending data from a sender  
3 of said computing environment over said transmission medium  
4 to said file system to be written to one or more storage  
5 media coupled to said file system.

1           18. The method of claim 17, wherein said means for  
2 sending comprises means for sending said data over said  
3 transmission medium using one or more buffers associated  
4 with said sender, and wherein said system further comprises  
5 means for receiving said data by said file system, said  
6 means for receiving comprising means for swapping one or  
7 more buffers associated with said file system with said one  
8 or more buffers associated with said sender.

1           19. The system of claim 18, wherein said means for  
2 swapping avoids copying data from said one or more buffers  
3 associated with said sender to said one or more buffers  
4 associated with said file system.

1           20. The system of claim 18, further comprising means  
2 for obtaining said one or more buffers associated with said  
3 file system, prior to said receiving.

1           21. The system of claim 18, further comprising means  
2 for returning said one or more buffers associated with said  
3 file system to free storage, subsequent to said swapping.

1           22. The system of claim 18, further comprising means  
2 for translating data in said one or more buffers associated  
3 with said file system to a format compatible with said file  
4 system, said translating being performed within said one or  
5 more buffers associated with said file system.

1           23. The system of claim 22, further comprising means  
2 for determining that translating is to be performed, said  
3 means for determining comprising means for checking a  
4 translation indicator to determine if translation is to be  
5 performed.

1           24. The system of claim 16, wherein said means for  
2 transmitting comprises means for sending data from said file  
3 system over said transmission medium to a receiver of said  
4 data.

1           25. The system of claim 24, wherein said means for  
2     sending comprises means for using a routine identified by  
3     said receiver to send said data, wherein said routine is  
4     provided one or more pointers to said data to be sent to  
5     said receiver.

1           26. The system of claim 24, wherein said means for  
2    sending comprises means for providing to said receiver one  
3    or more pointers to said data.

1           27. A system of reducing data movement within a  
2 computing environment, said system comprising:

3           a sender of said computing environment adapted to  
4 send data to a file system of said computing  
5 environment, wherein the sending comprises using one or  
6 more buffers associated with said sender; and

7           said file system adapted to receive said data,  
8 wherein the receiving comprises swapping one or more  
9 buffers associated with said file system with said one  
10 or more buffers associated with said sender.

1           28. The system of claim 27, wherein the swapping  
2 avoids copying data from said one or more buffers associated  
3 with said sender to said one or more buffers associated with  
4 said file system.

1           29. A system of translating data from one format to  
2 another format, said system comprising:

3           means for determining that data located in at  
4 least one buffer associated with a file system usable  
5 in writing data to one or more storage media coupled to  
6 said file system is to be translated from one format to  
7 another format; and

8           means for translating at least a portion of said  
9 data in said at least one buffer, said translating  
10 being performed within said at least one buffer  
11 associated with said file system without requiring  
12 copying of said at least portion of said data to one or  
13 more other buffers.

1           30. The system of claim 29, wherein said means for  
2 determining comprises means for checking a translation  
3 indicator to determine if translation is to be performed.

1           31. At least one program storage device readable by a  
2 machine, tangibly embodying at least one program of  
3 instructions executable by the machine to perform a method  
4 of reducing data movement within a computing environment,  
5 said method comprising:

6           transmitting data between a file system of a  
7 computing unit of said computing environment and a  
8 transmission medium of said computing environment; and

9           wherein said transmitting comprises bypassing non-  
10 file system buffers of said computing unit in  
11 performing the transmission.

1           32. The at least one program storage device of claim  
2 31, wherein said transmitting comprises sending data from a  
3 sender of said computing environment over said transmission  
4 medium to said file system to be written to one or more  
5 storage media coupled to said file system.

1           33. The at least one program storage device of claim  
2 32, wherein said sending comprises sending said data over  
3 said transmission medium using one or more buffers  
4 associated with said sender, and wherein said method further  
5 comprises receiving said data by said file system, said  
6 receiving comprising swapping one or more buffers associated  
7 with said file system with said one or more buffers  
8 associated with said sender.



1        34. The at least one program storage device of claim  
2        33, wherein said swapping avoids copying data from said one  
3        or more buffers associated with said sender to said one or  
4        more buffers associated with said file system.

1        35. The at least one program storage device of claim  
2        33, wherein said method further comprises obtaining said one  
3        or more buffers associated with said file system, prior to  
4        the receiving.

1        36. The at least one program storage device of claim  
2        33, wherein said method further comprises returning said one  
3        or more buffers associated with said file system to free  
4        storage, subsequent to the swapping.

1        37. The at least one program storage device of claim  
2        33, wherein said method further comprises translating data  
3        in said one or more buffers associated with said file system  
4        to a format compatible with said file system, said  
5        translating being performed within said one or more buffers  
6        associated with said file system.

1        38. The at least one program storage device of claim  
2        37, wherein said method further comprises determining that  
3        translating is to be performed, said determining comprising  
4        checking a translation indicator to determine if translation  
5        is to be performed.

1           39. The at least one program storage device of claim  
2 31, wherein said transmitting comprises sending data from  
3 said file system over said transmission medium to a receiver  
4 of said data.

1           40. The at least one program storage device of claim  
2 39, wherein said sending comprises using a routine  
3 identified by said receiver to send said data, wherein said  
4 routine is provided one or more pointers to said data to be  
5 sent to said receiver.

1           41. The at least one program storage device of claim  
2 39, wherein said sending comprises providing to said  
3 receiver one or more pointers to said data.

1 42. An article of manufacture, comprising:

2 at least one computer usable medium having  
3 computer readable program code means embodied therein  
4 for causing the reducing of data movement within a  
5 computing environment, the computer readable program  
6 code means in said article of manufacture comprising:

7 computer readable program code means for  
8 causing a computer to send data from a sender of  
9 said computing environment to a file system of  
10 said computing environment, said computer readable  
11 program code means for causing a computer to send  
12 comprising computer readable program code means  
13 for causing a computer to use one or more buffers  
14 associated with said sender; and

15 computer readable program code means for  
16 causing a computer to receive by said file system  
17 the data, wherein said computer readable program  
18 code means for causing a computer to receive  
19 comprises computer readable program code means for  
20 causing a computer to swap one or more buffers  
21 associated with said file system with said one or  
22 more buffers associated with said sender.

1           43. The article of manufacture of claim 42, wherein  
2   said computer readable program code means for causing a  
3   computer to swap avoids copying data from said one or more  
4   buffers associated with said sender to said one or more  
5   buffers associated with said file system.

43. The article of manufacture of claim 42, wherein  
said computer readable program code means for causing a  
computer to swap avoids copying data from said one or more  
buffers associated with said sender to said one or more  
buffers associated with said file system.

1           44. At least one program storage device readable by a  
2 machine, tangibly embodying at least one program of  
3 instructions executable by the machine to perform a method  
4 of translating data from one format to another format, said  
5 method comprising:

6           determining that data located in at least one  
7 buffer associated with a file system usable in writing  
8 data to one or more storage media coupled to said file  
9 system is to be translated from one format to another  
10 format; and

11           translating at least a portion of said data in  
12 said at least one buffer, said translating being  
13 performed within said at least one buffer associated  
14 with said file system without requiring copying of said  
15 at least portion of said data to one or more other  
16 buffers.

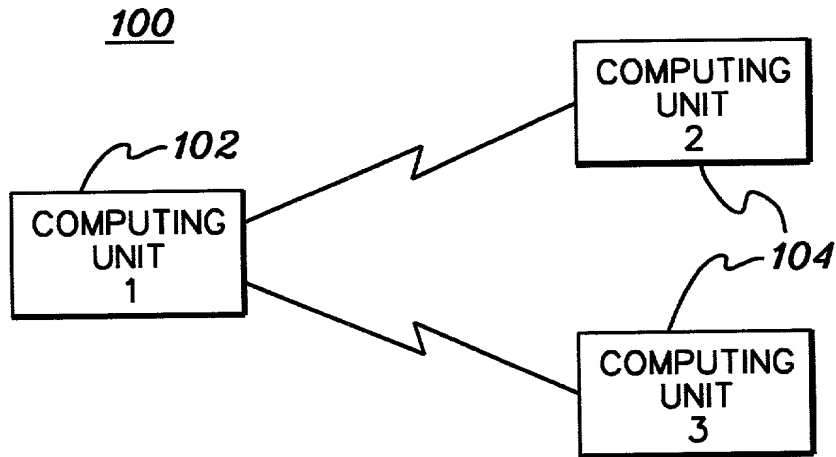
1           45. The at least one program storage device of claim  
2 44, wherein said determining comprises checking a  
3 translation indicator to determine if translation is to be  
4 performed.

\* \* \* \* \*

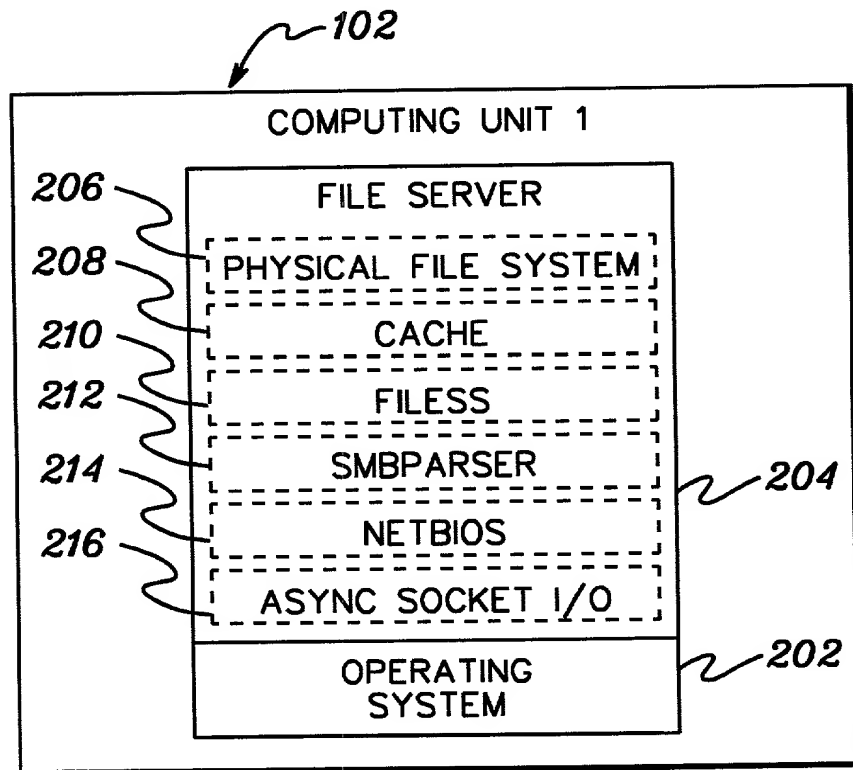
**METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING  
DATA MOVEMENT WITHIN A COMPUTING ENVIRONMENT**

**Abstract of the Disclosure**

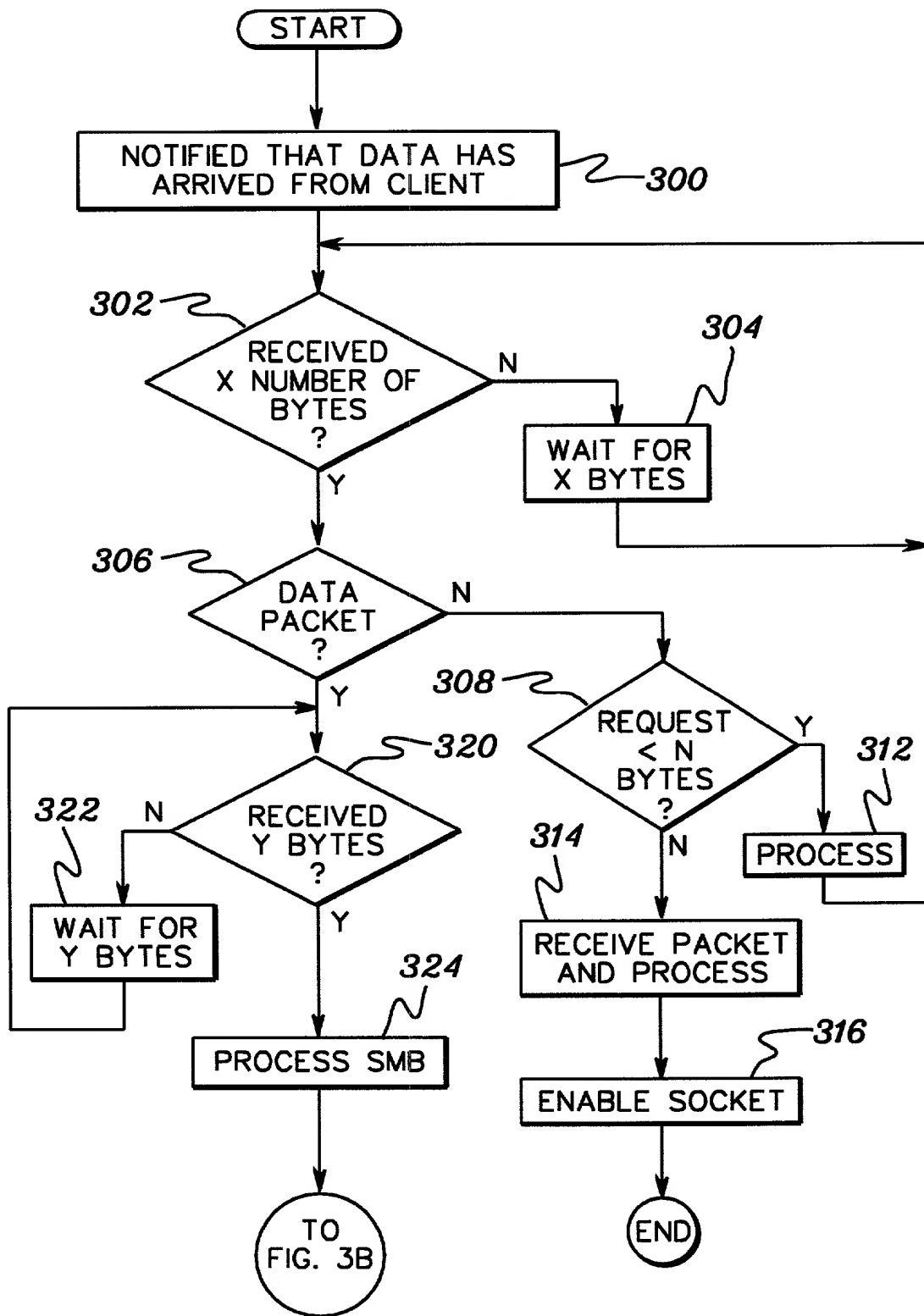
Data movement within a computing environment is at the  
5 very least reduced. Data is transmitted between a file  
system of the computing environment and a transmission  
medium of that environment. The transmission includes  
bypassing non-file system buffers in performing the  
transmission. For example, when data is sent to the file  
10 system to be written to one or more storage media, the file  
system swaps one or more buffers of the file system with the  
one or more buffers containing the data. The swapping does  
not require the copying of data. Further, for a read  
operation, the file system calls a routine, which is  
15 provided with one or more pointers to the data that is to be  
sent to a requester of the data.



*fig. 1*



*fig. 2*



*fig. 3A*



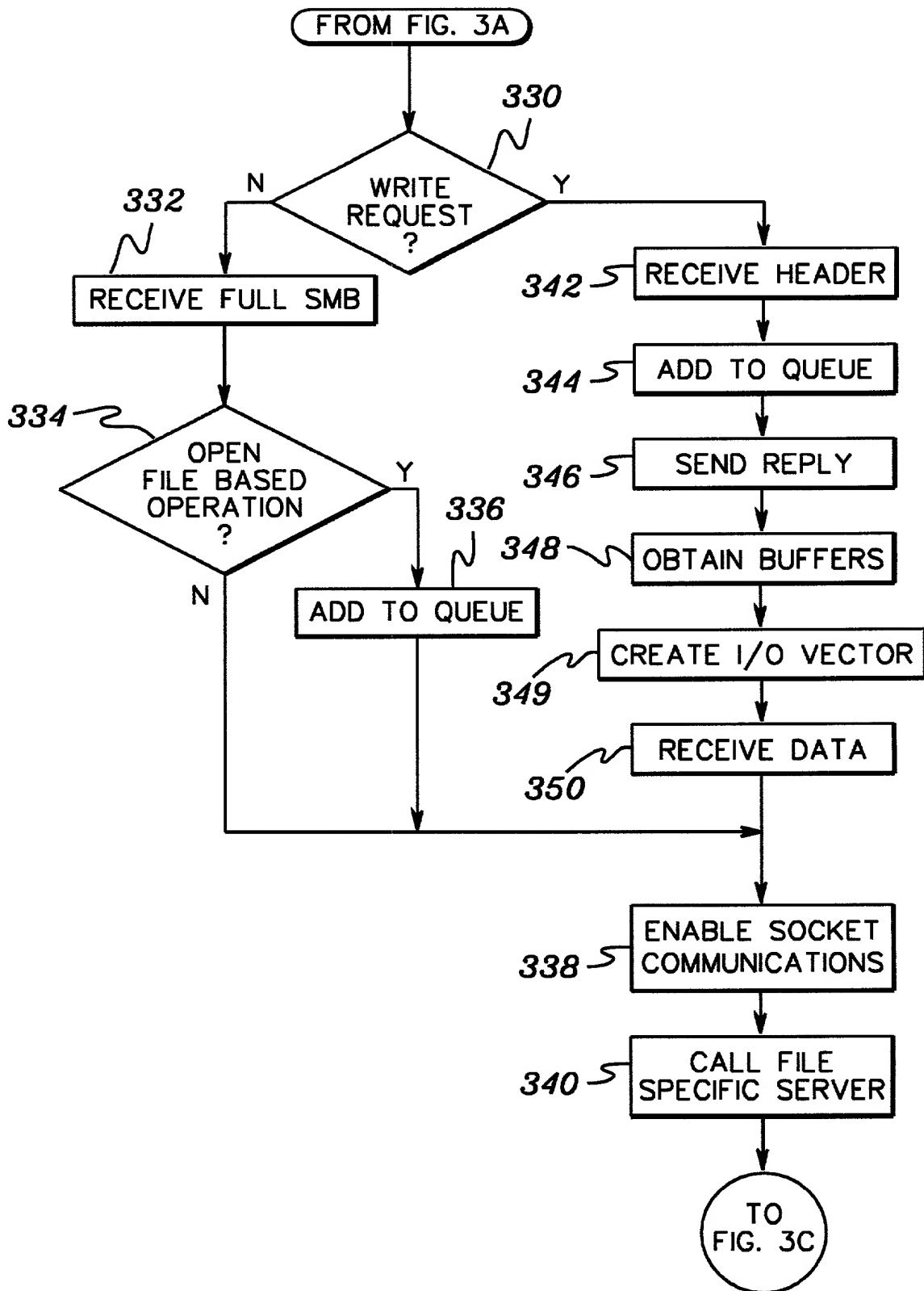


fig. 3B

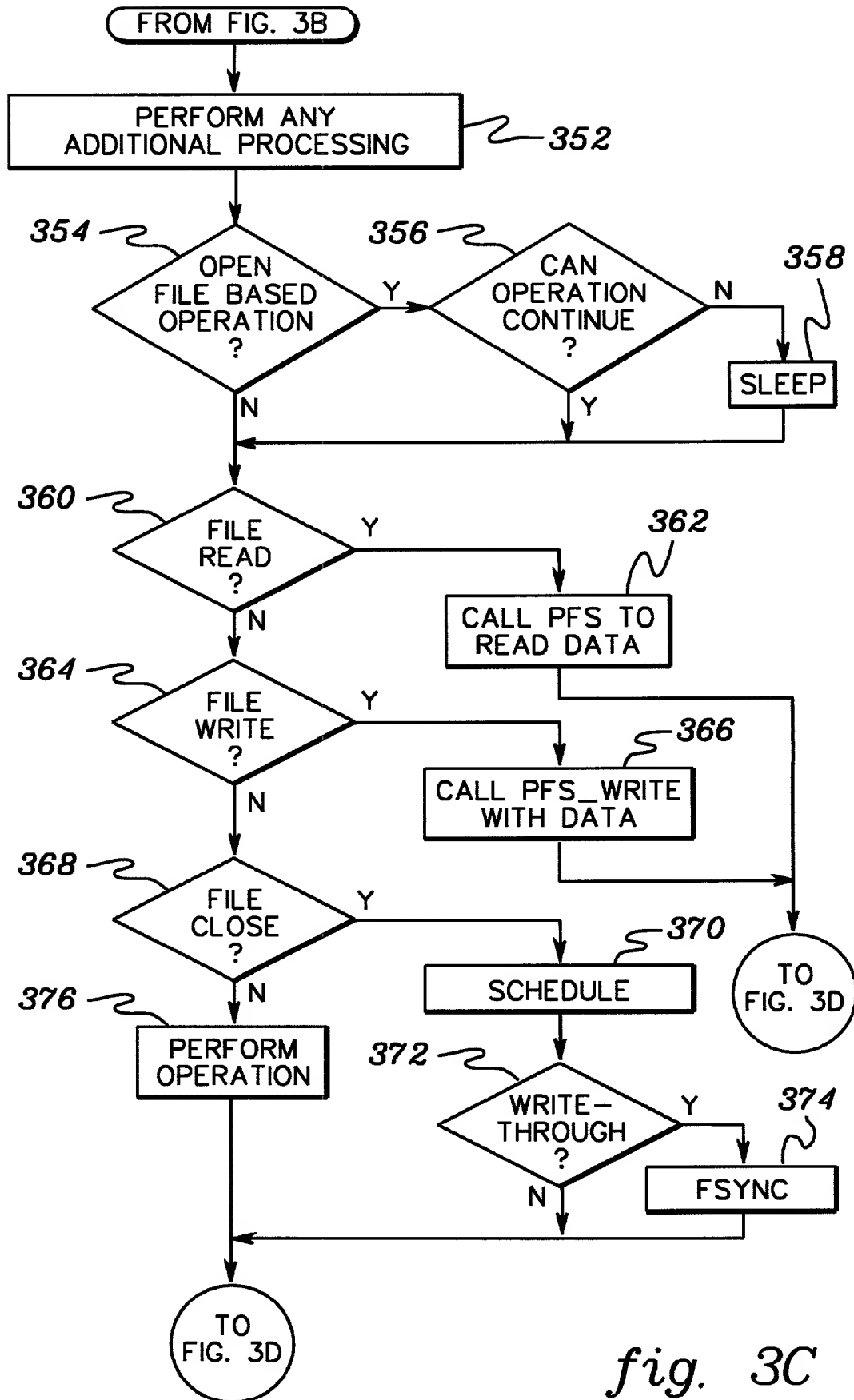
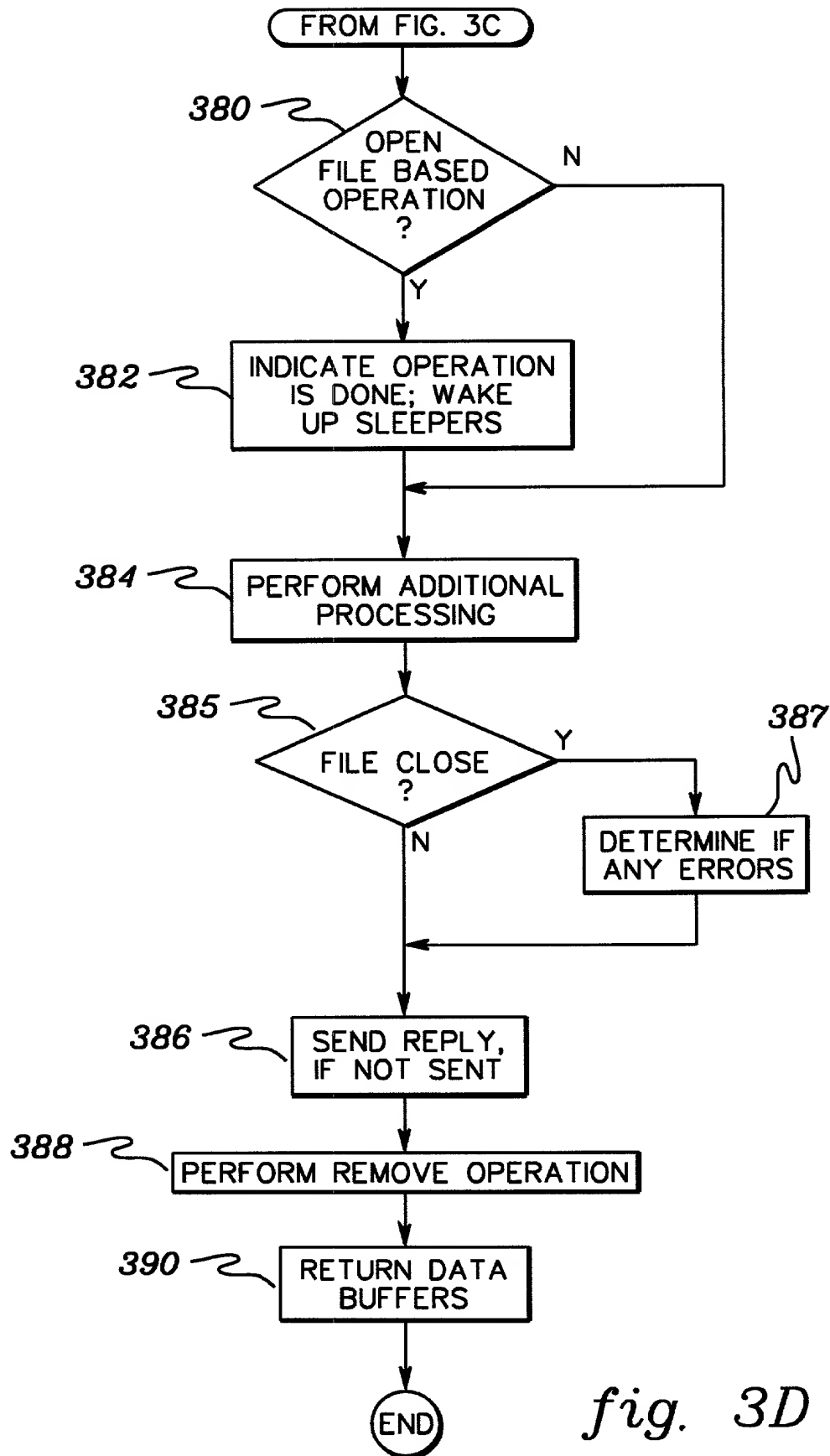


fig. 3C



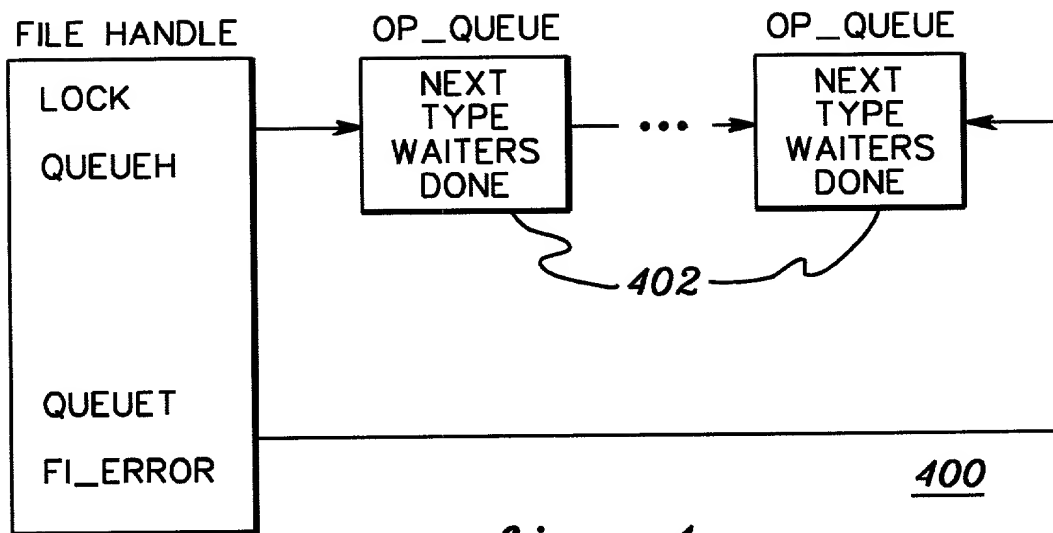


fig. 4

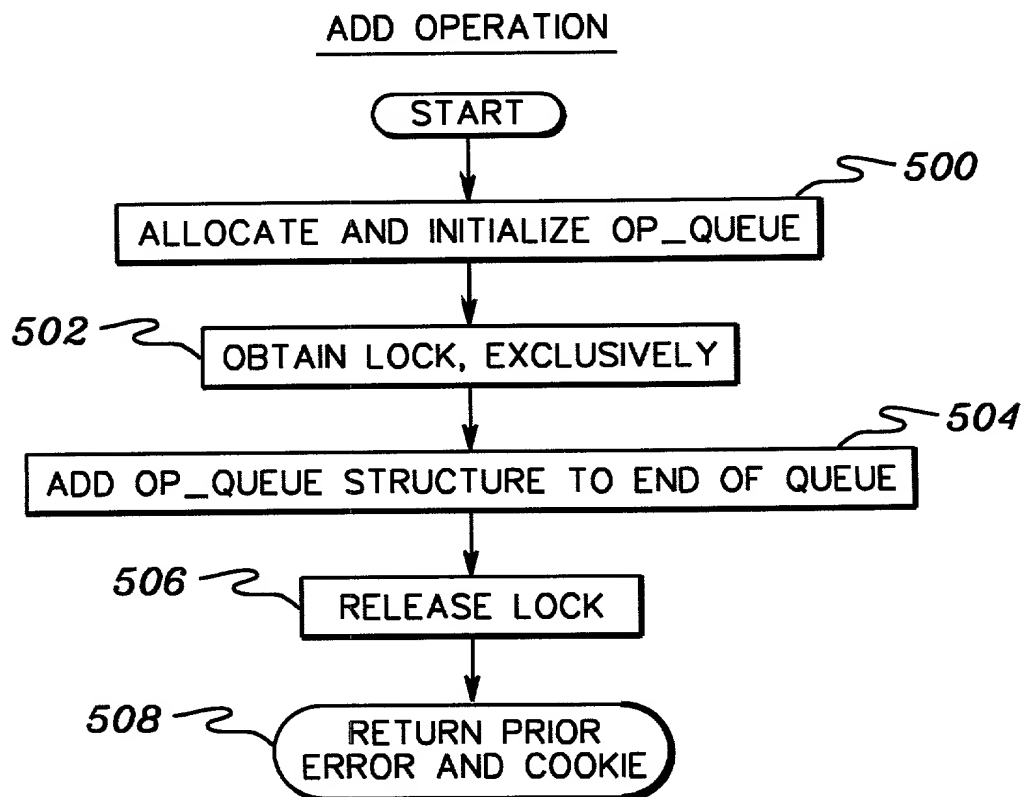
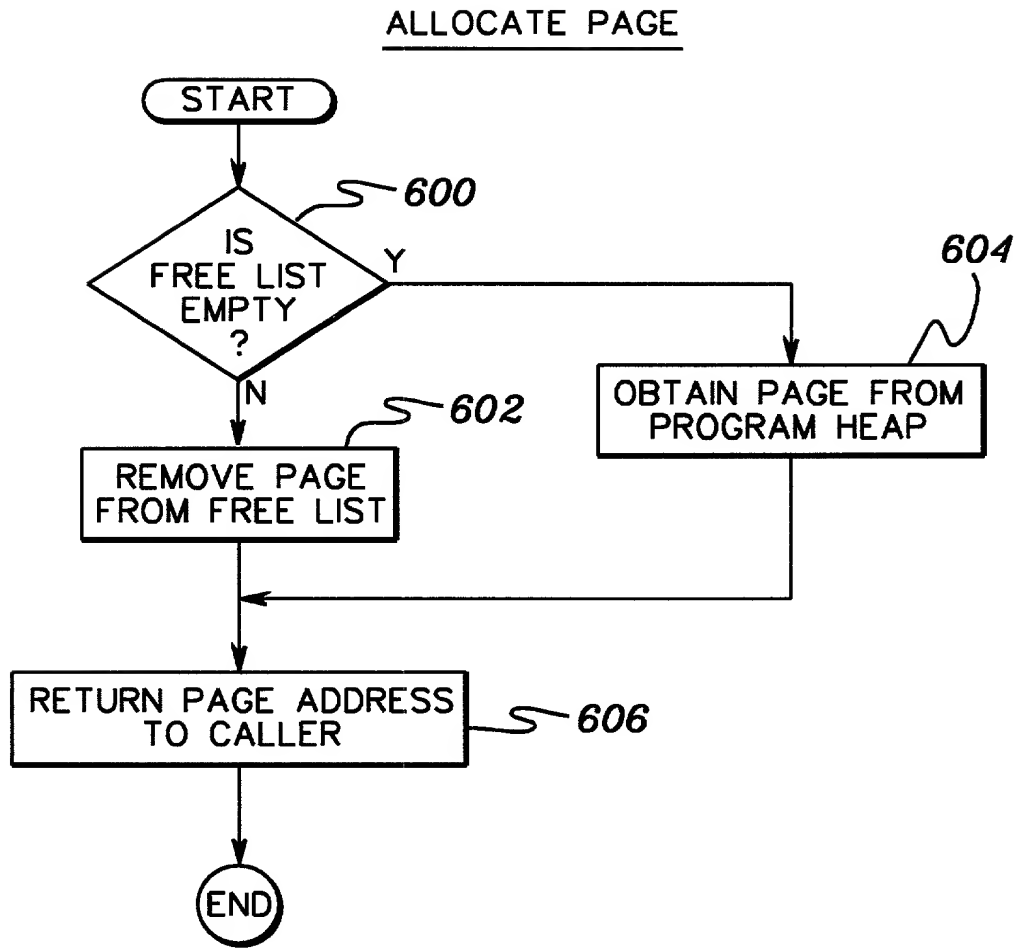
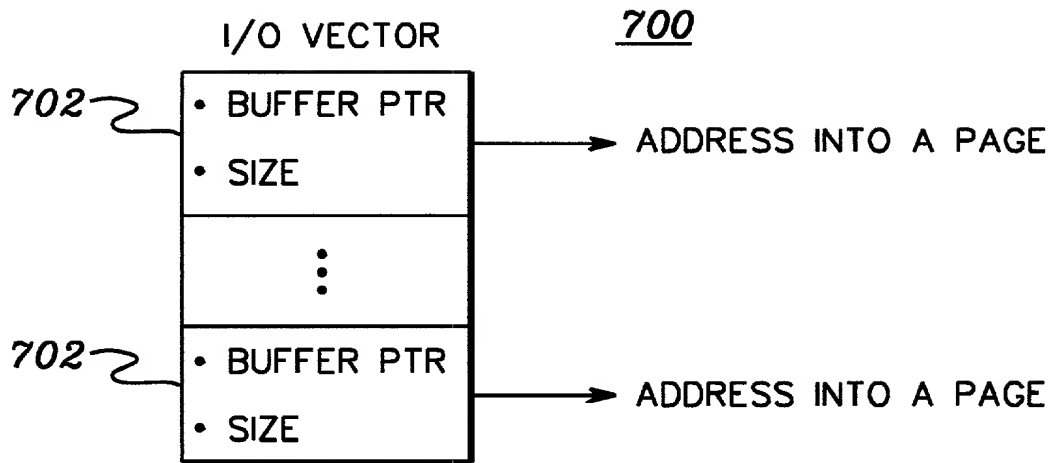


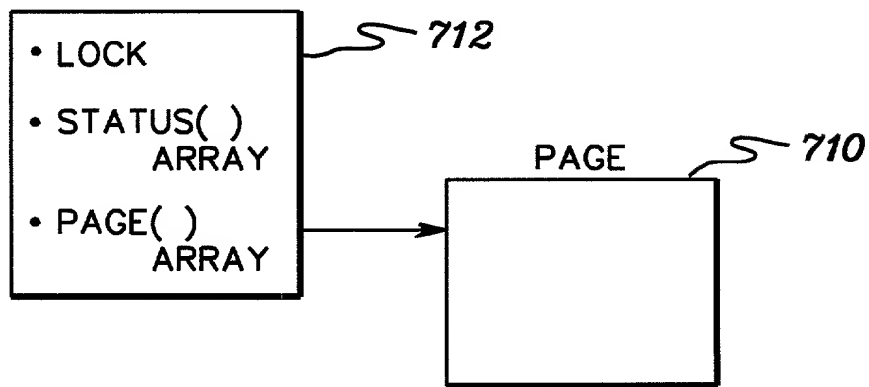
fig. 5



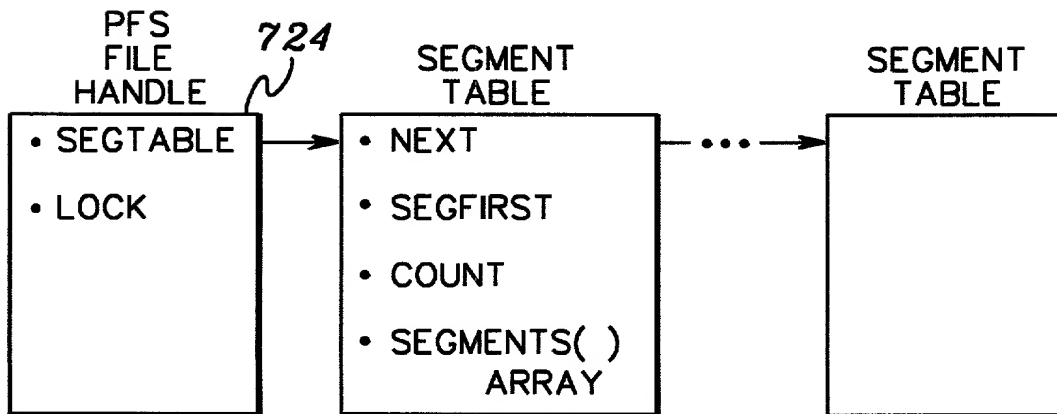
*fig. 6*



*fig. 7A*



*fig. 7B*



*fig. 7C*

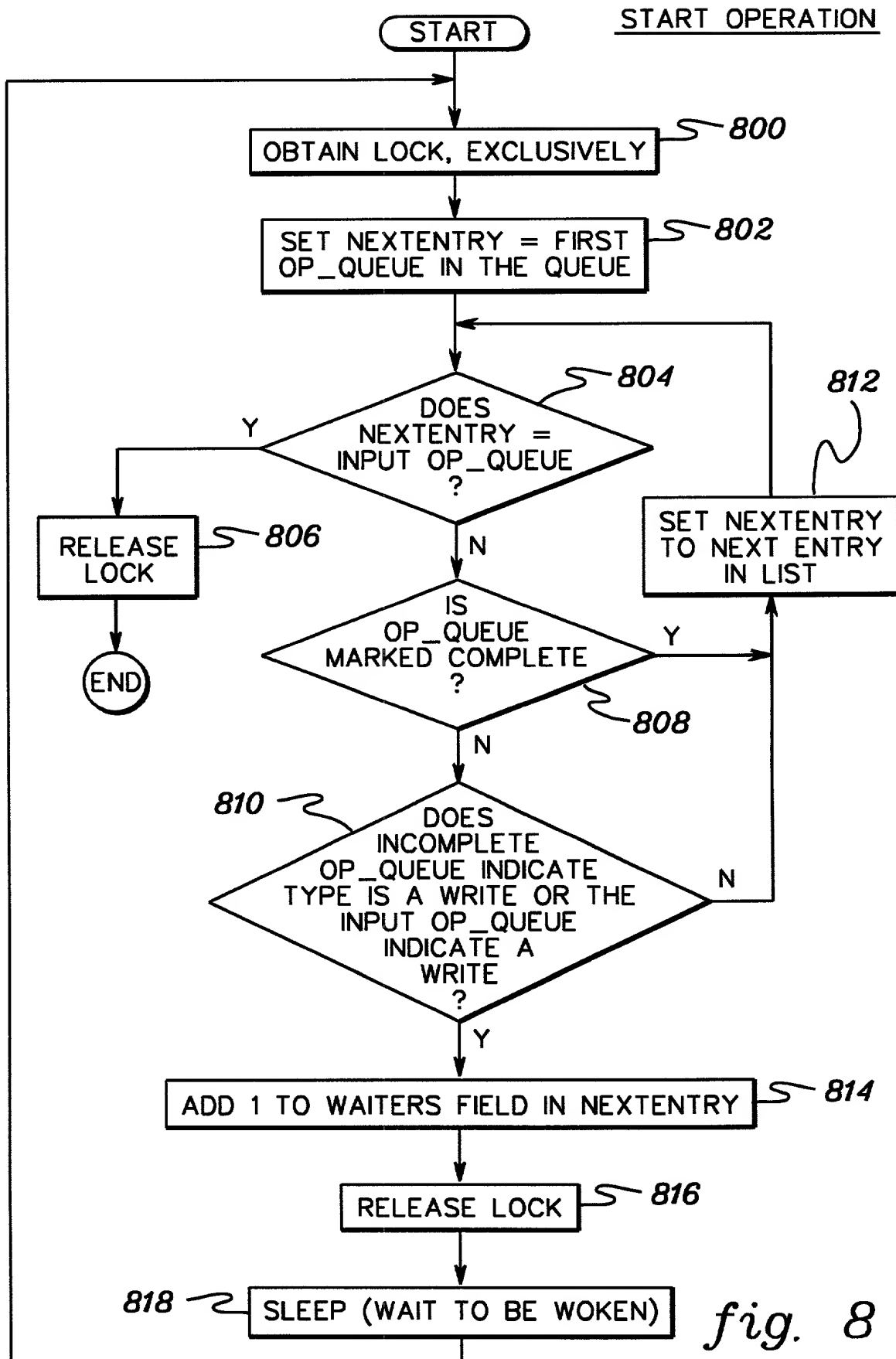
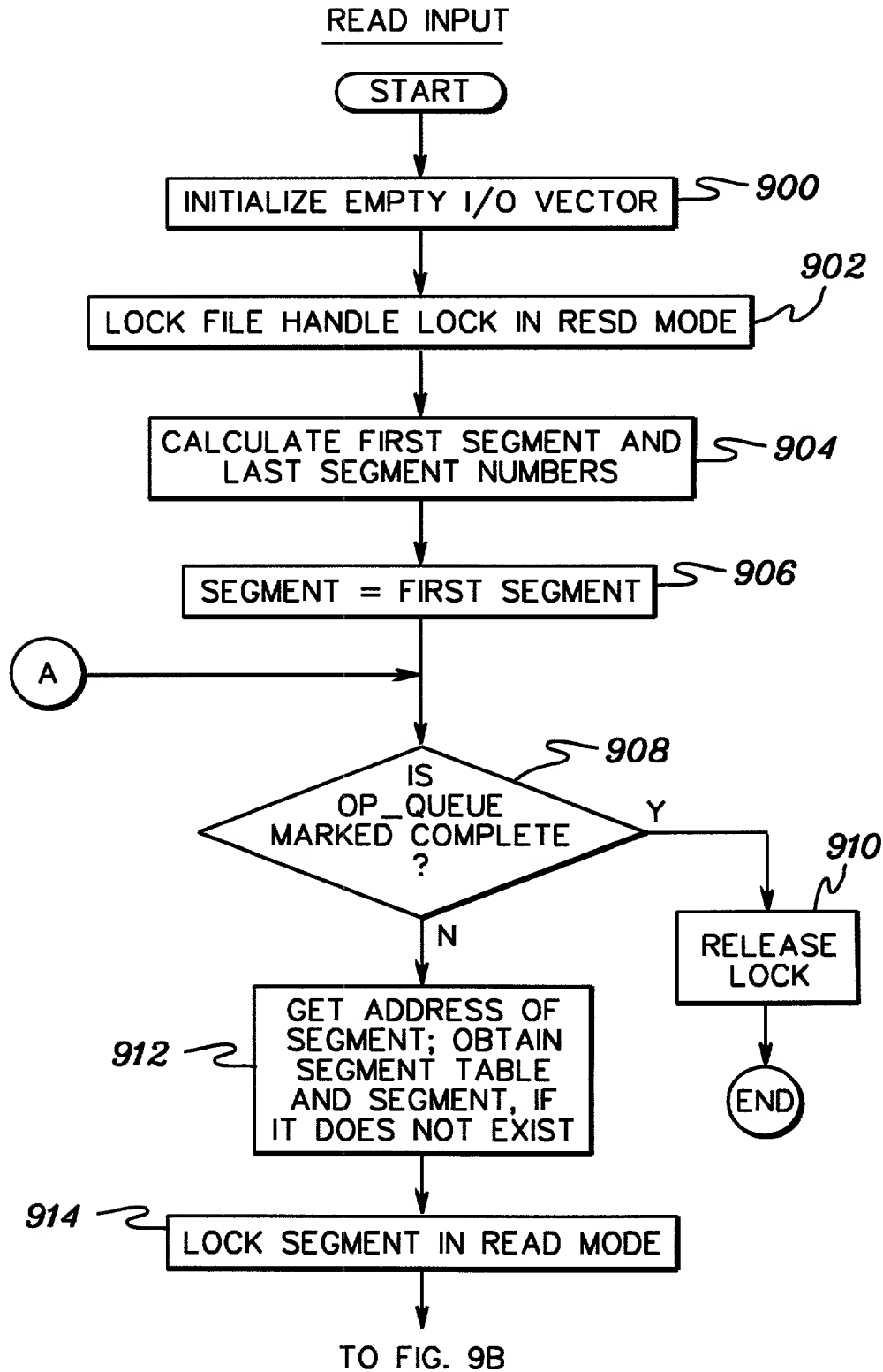
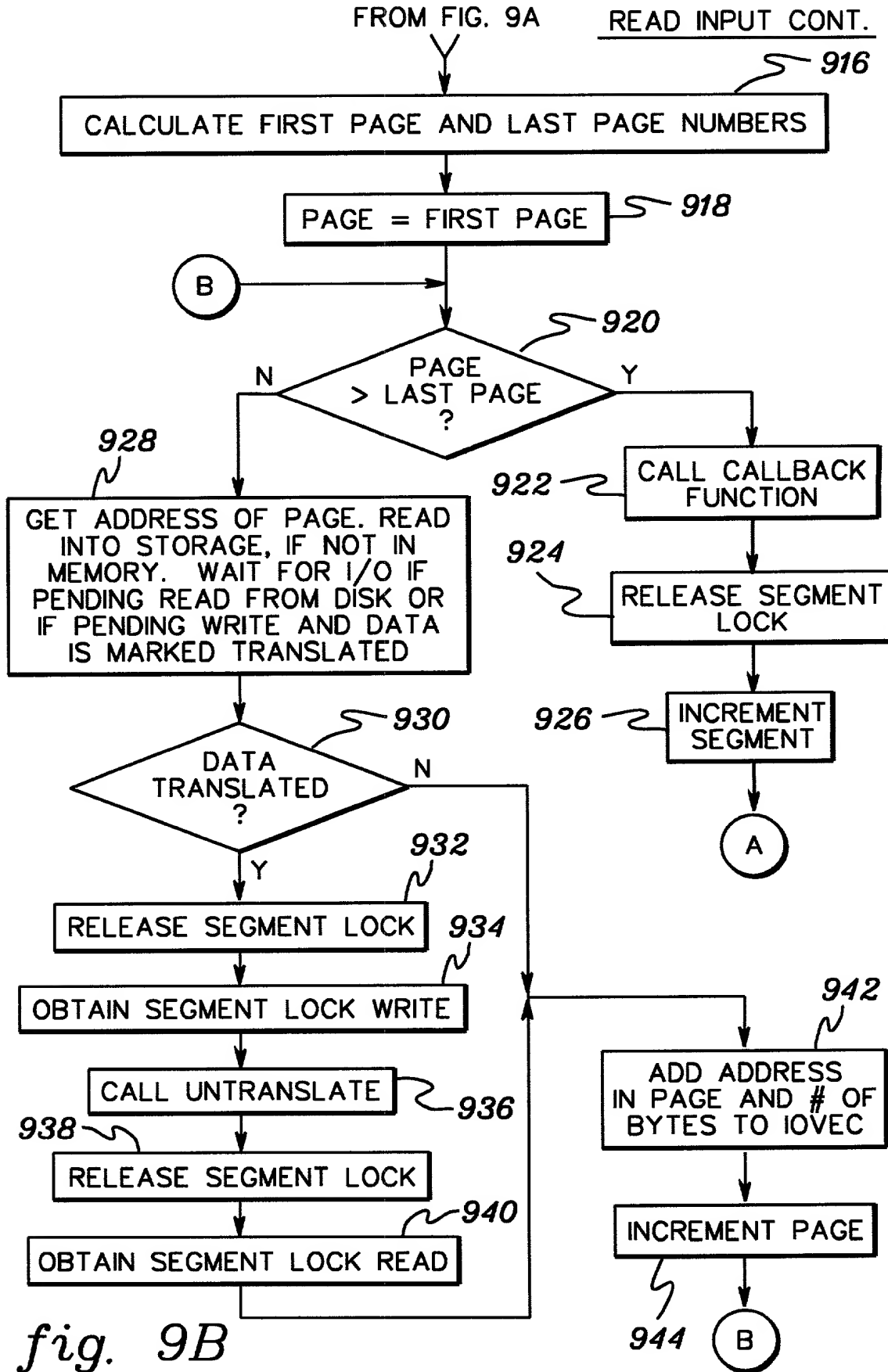


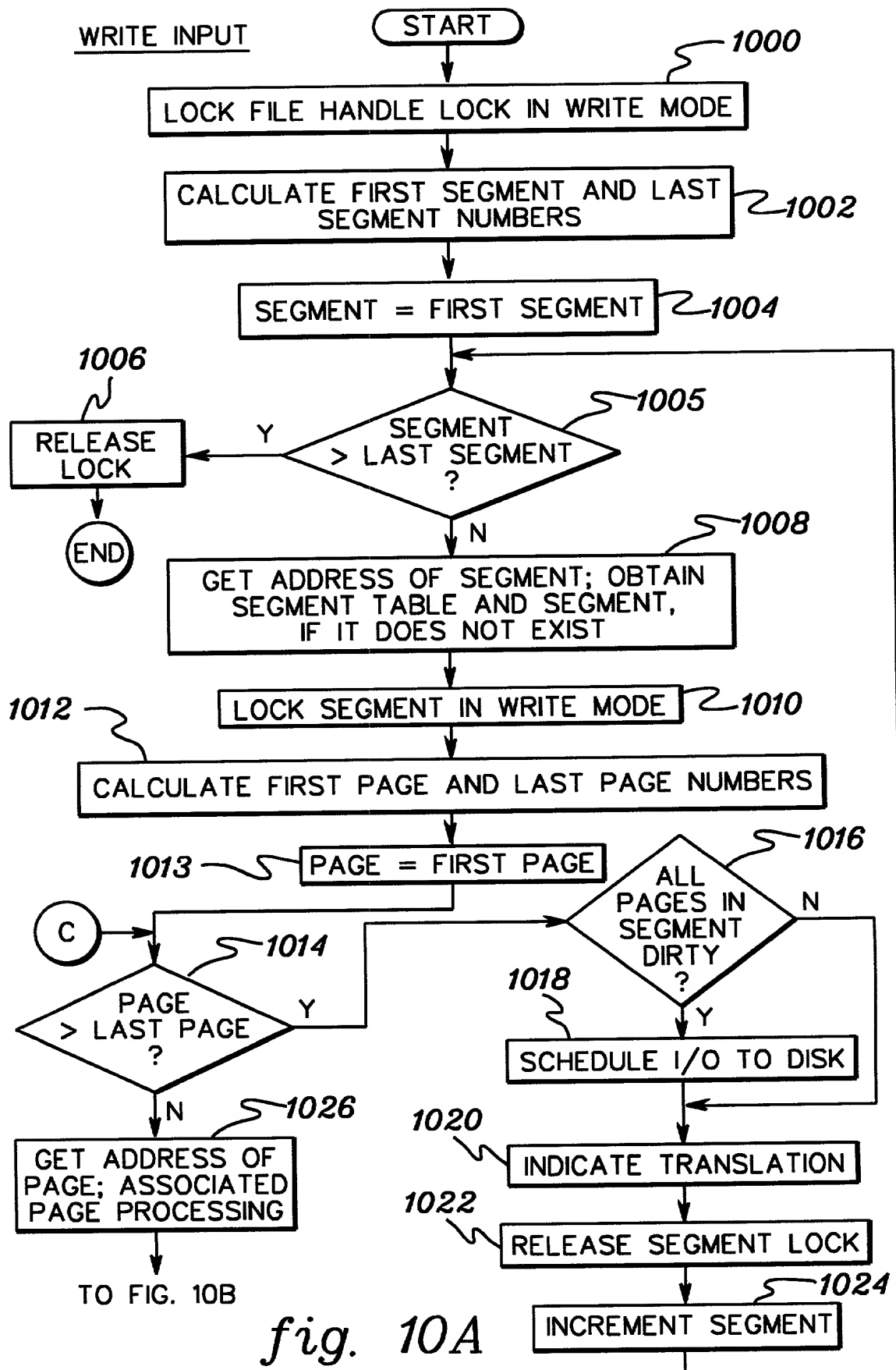
fig. 8



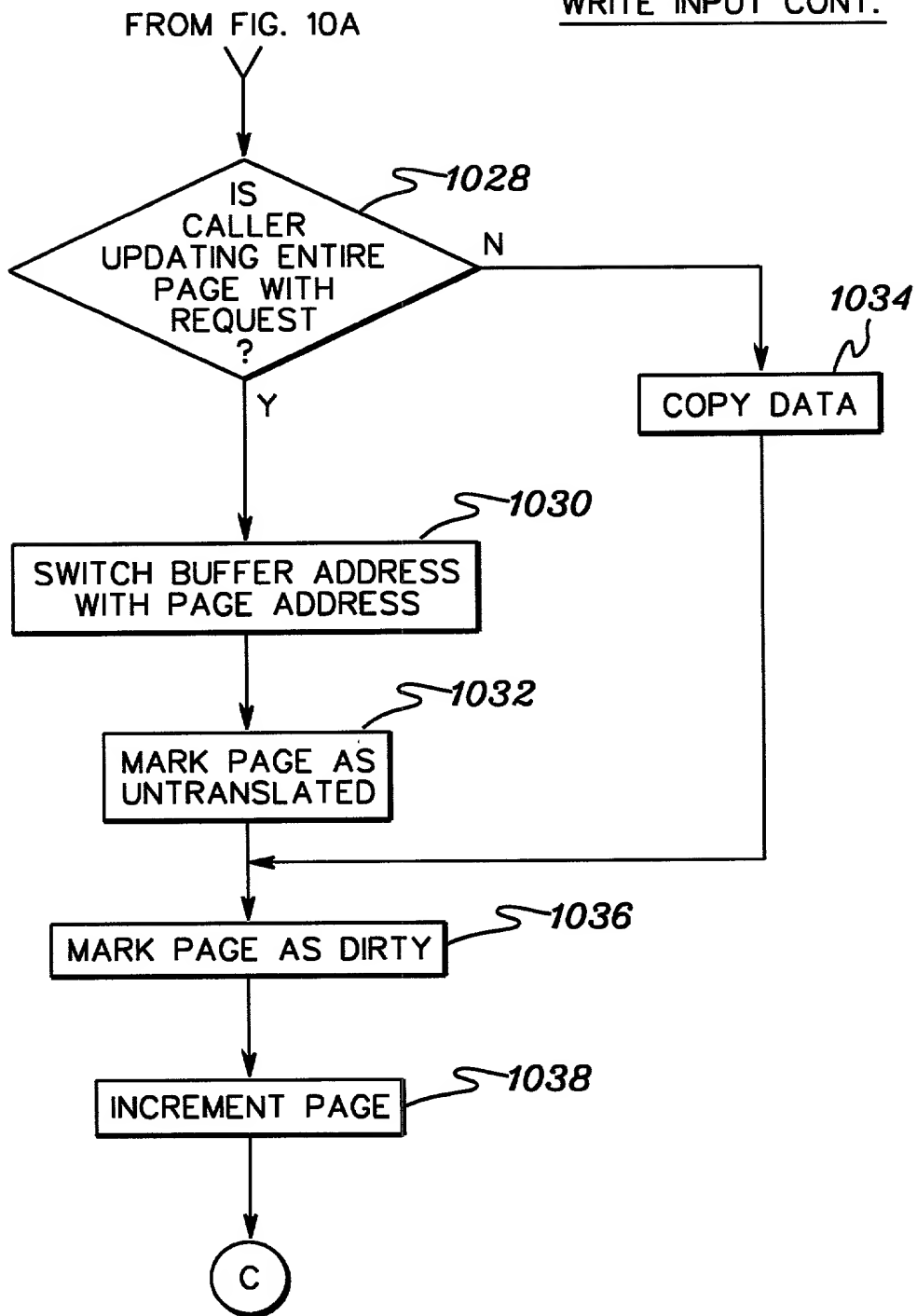
*fig. 9A*





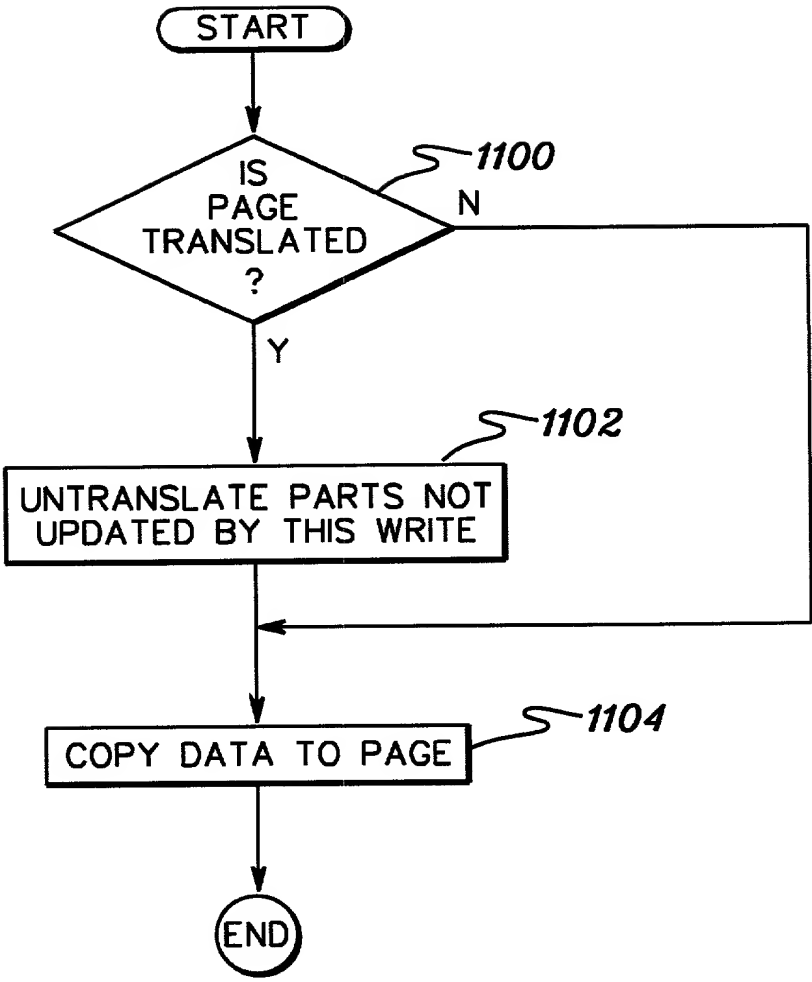


WRITE INPUT CONT.



*fig. 10B*

COPY DATA



*fig. 11*

STOP OPERATION

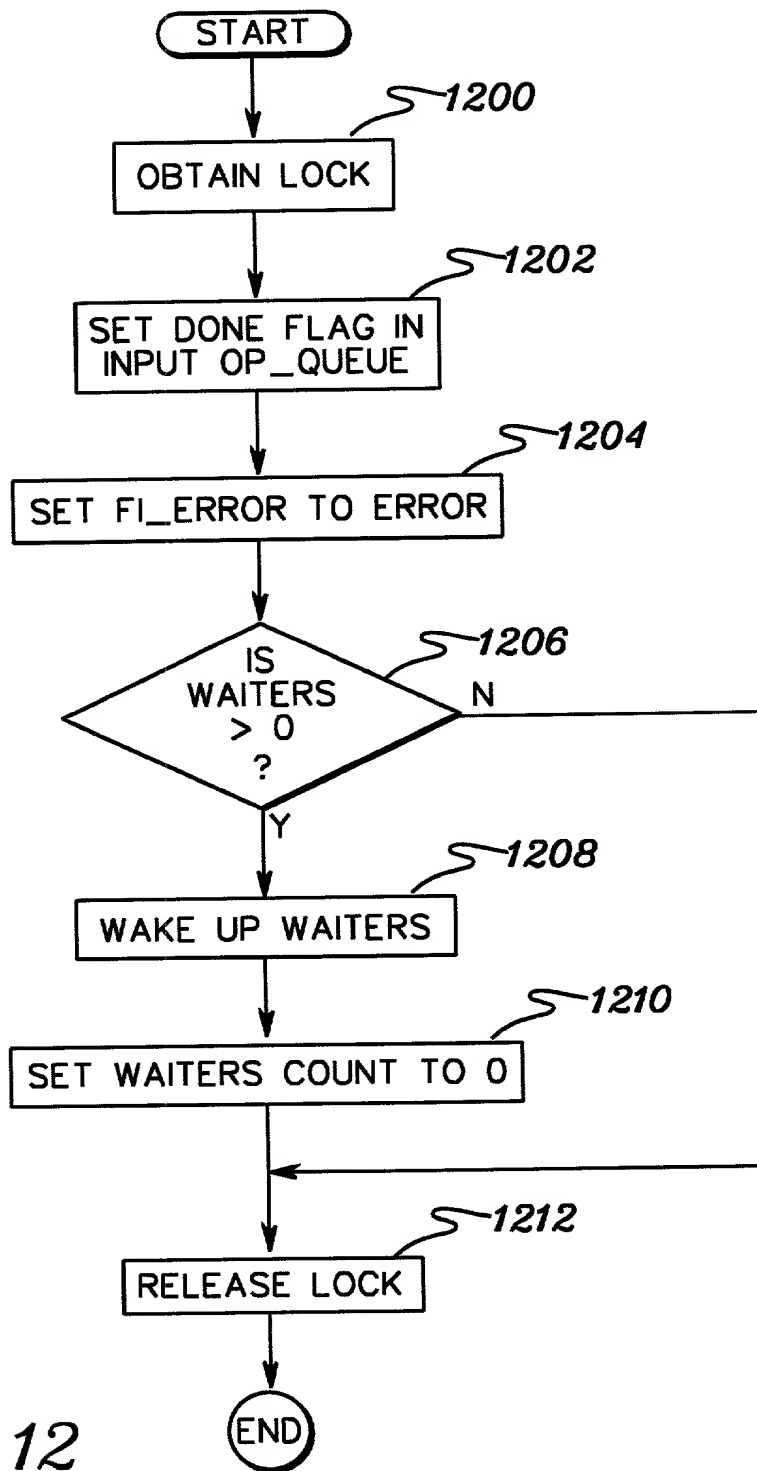


fig. 12

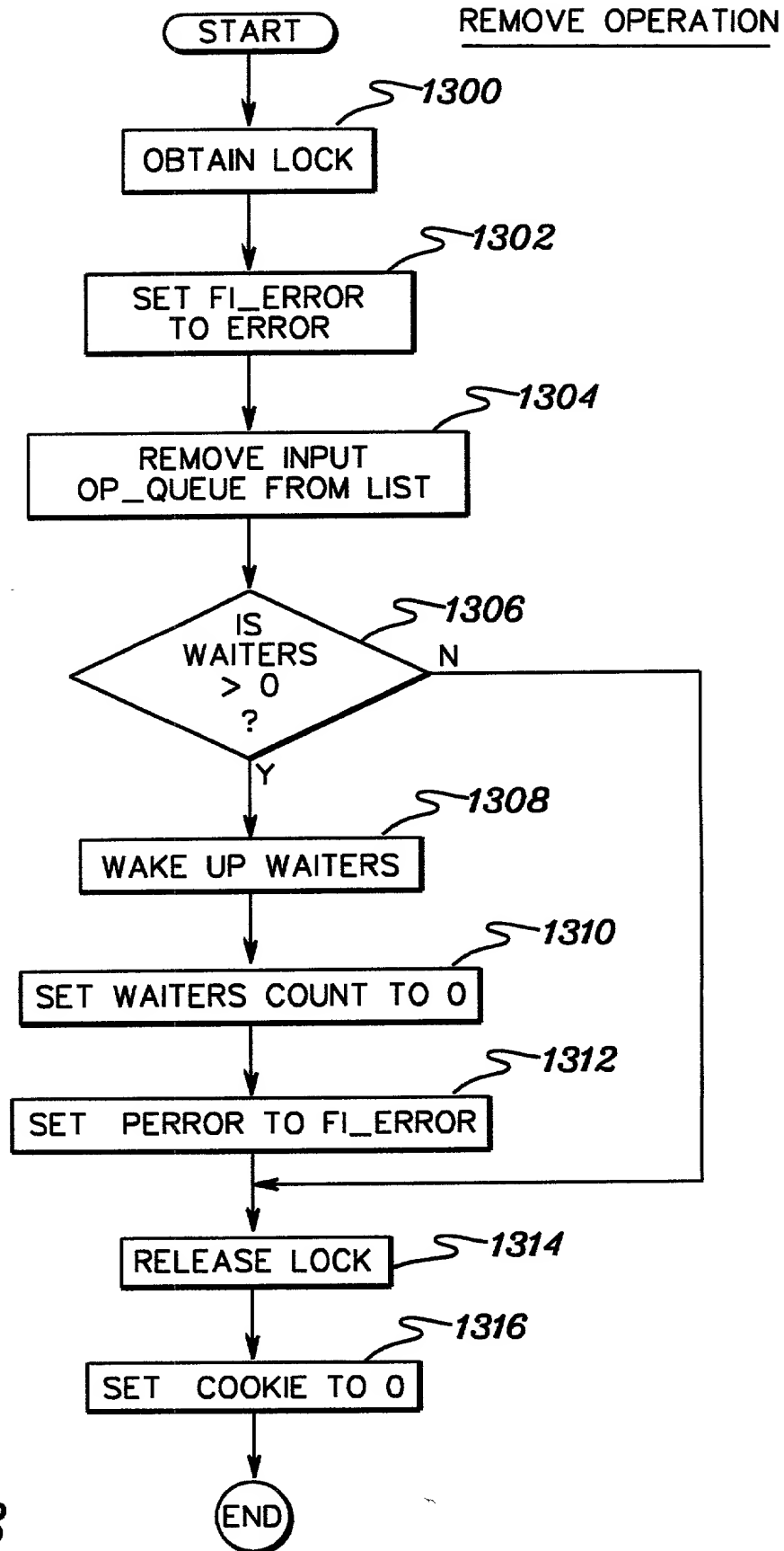
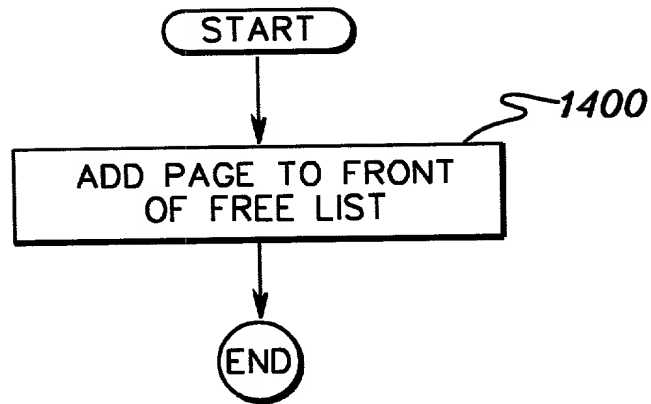


fig. 13

FREE PAGE



*fig. 14*

## DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name. I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

**METHOD, SYSTEM AND PROGRAM PRODUCTS FOR REDUCING  
DATA MOVEMENT WITHIN A COMPUTING ENVIRONMENT**

the specification of which (check one)

  X   is attached hereto.

\_\_\_\_\_ was filed on \_\_\_\_\_ as United States  
Application Number or PCT International Application  
Number \_\_\_\_\_ and was amended on  
\_\_\_\_\_.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above. I acknowledge the duty to disclose to the U.S. Patent and Trademark Office all information known to me to be material to patentability as defined in 37 CFR § 1.56.

I hereby claim foreign priority benefits under 35 U.S.C. § 119(a)-(d) or § 365(b) of any foreign application(s) for patent or inventor's certificate, or § 365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below any foreign application for patent or inventor's certificate, or PCT International application having a filing date before that of the application on which priority is claimed.

**Priority Claimed**

_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	____ Yes	____ No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	____ Yes	____ No

I hereby claim the benefit under 35 U.S.C. §119(e) of any United States provisional application(s) listed below.

_____ (Application Number)	_____ (Filing Date)
_____ (Application Number)	_____ (Filing Date)



I hereby claim the benefit under 35 U.S.C. § 120 of any United States application(s), or § 365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose material information as defined in 37 CFR §1.56(a) which occurred between the filing date of the prior application and the national or PCT International filing date of this application:

(Appl. Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)
--------------------	---------------	---

(Appl. Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)
--------------------	---------------	---

**POWER OF ATTORNEY:** As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith:

Christopher A. Hughes, Esq.	Reg. No. 26,914
Edward A. Pennington, Esq.	Reg. No. 32,588
John E. Hoel, Esq.	Reg. No. 26,279
Joseph C. Redmond, Jr., Esq.	Reg. No. 18,753
David L. Adour, Esq.	Reg. No. 29,604
Lawrence R. Fraley, Esq.	Reg. No. 26,885
Arthur J. Samodovitz, Esq.	Reg. No. 31,297
William H. Steinberg, Esq.	Reg. No. 28,540
John R. Pivnichny, Ph.D.	Reg. No. 43,001
Jeff Rothenberg, Esq.	Reg. No. 26,429
Kevin P. Radigan, Esq.	Reg. No. 31,789
Blanche E. Schiller, Esq.	Reg. No. 35,670

**Send Correspondence to:**

Blanche E. Schiller, Esq.  
 HESLIN & ROTHENBERG, P.C.  
 5 Columbia Circle  
 Albany, New York 12203-5160  
 Telephone: (518) 452-5600  
 Facsimile: (518) 452-5579

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

ADDED PAGE(S) TO COMBINED DECLARATION AND POWER OF ATTORNEY  
FOR SIGNATURE BY FIRST AND SUBSEQUENT INVENTORS

Full Name of first joint inventor: Scott Thomas Marcotte

Signature: Scott Thomas Marcotte Date: 11/17/99

Residence: 337 Third Avenue, Vestal, New York 13850

Citizenship: United States of America

Post Office Address: 337 Third Avenue, Vestal, New York 13850

Full Name of second joint inventor:

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Residence:

Citizenship:

Post Office Address:

Full Name of third joint inventor:

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Residence:

Citizenship:

Post Office Address: